UNIVERSITÀ DI PISA
Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
Ingegneria dell'Informazione

Tesi di Dottorato di Ricerca

Language Resource Infrastructure(s)

Riccardo Del Gratta

Anno 2011

Università di Pisa

Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
Ingegneria dell'Informazione

Ph.D. Thesis

Language Resource Infrastructure(s)

Candidate:

Riccardo Del Gratta

Supervisors:

Prof. Ing. Luca Simoncini

Dr. Nicoletta Calzolari

Dr. Alessandro Enea

2011

SSD ING-INF/05

Sommario

Non esiste una sola Infrastruttura di Risorse Linguistiche, ma molte infrastrutture e tutte tra loro diverse, anche se con aspetti comuni. Il motivo del plurale, la (s), nel titolo della tesi è esattamente questo.
La comunità dei linguisti è molto variegata: studiosi di scienze sociali ed umane sono linguisti, come linguisti sono quelli che direttamente si occupano di (o forniscono consulenze in) ambiti molto più tecnici come la traduzione automatica, l'estrazione di informazioni da testi, il *question-answering* fino ai motori di ricerca presenti sul Web. Ogni sotto comunità linguistica ha le proprie esigenze da richiedere ad una Infrastruttura di Risorse Linguistiche: disponibilità di risorse, possibilità di scaricare liberamente software normalmente a pagamento, presenza di commenti e valutazioni sulle risorse disponibili ed ancora altro. Possiamo affermare che, spesso, sono i requisiti utenti a guidare il design architetturale ed il modello delle infrastrutture, mentre le tecnologie più prettamente informatiche sono usate per trovare soluzioni a tali requisiti. A conferma di questo aspetto, possiamo citare due progetti europei, METANET e PANACEA: il primo è volto alla creazione di un network di repository di tool e dati languistici accessibili da una più ampia comunità di linguisti, mentre il secondo è una piattaforma volta alla creazione di un network di risorse linguistiche in ambito multilingue e della Machine Translation, pensato per essere usato da industrie in tali ambiti.
Entrambi i progetti hanno la comunità dei linguisti come promotori (*provider* di servizi linguistici) ma diverse comunità di utenti esterni a cui i servizi sono rivolti (*consumer*).
METANET ha come *consumer* ancora la comunità dei linguisti computazionali, mentre PANACEA ha la comunità di industrie legate alla Machine Translation come comunità *consumer*. La diversità degli utenti finali porta a diversi requisiti utente e, quindi, a caratteristiche differenti nelle infrastrutture.
In questa tesi descriviamo sia gli aspetti comuni che specifici delle Infrastrutture di Risorse Linguistiche e mettiamo in risalto il nostro apporto alla progettazione ad alto livello delle infrastrutture di entrambi i progetti. Nello specifico riportiamo i nostri contributi nell'ambito della definizione dei moduli architetturali connessi alla autenticazione ed autorizzazione, e più in generale alla gestione degli utenti, ed al loro accesso alle risorse linguistiche.

Abstract

We have added an "(s)" to the title of this thesis because there is not a single one "Language Resource Infrastructure" but many Language Resource Infrastructures. In fact, the language resource infrastructures are all partially alike, since they have many common aspects, but every single language resource infrastructure is peculiar in its own way, since it has its own distinguishing characteristics.
The community of linguists is very wide-ranging: human and social science scientists are linguists, as linguists are those who work in more technical environments such as Machine Translation, Information Extraction, Question-Answering, search engines and technologies available on the Web. Each sub community wants that the Language Resource Infrastructures will address its own requirements: resource availability, free download of resources normally available for-fee, feedback, comments on language resources, evaluation of language resources and so on. We can say that user requirements drive the designing and modeling of the infrastructures more than information technology, whose experts are asked to solve issues and provide solution for the user requirements. To confirm this aspect, we can cite two European projects, METANET and PANACEA: the former aims at building a network of repositories of language resources and technologies widely available for an increasing linguistic community, while the latter is a platform designed for the lexical acquisition and managing multilingualism and Machine Translation issues for small and medium enterprises focused on such topics.
Both projects have the language resource community as internal users, that is to say, as *providers* of language services, but a different target with respect to the *consumers* of language resources and services.
METANET is a project made by computational linguists for (computational) linguists, while PANACEA provides services for the Machine Translation industrial community. As a consequence, different requirements have led to different language resource infrastructures.
In this thesis we describe both common and specific aspects of Language Resource Infrastructures and point out our contribution to the modeling of the high level architecture of the infrastructure in both projects. In particular, we report our contribution in the area of Access and Identity Management, specifically in the user management and his/her access to language resources

. . . ai miei genitori . . .

Contents

Chapter 1

Introduction

Different initiatives, both in and outside Europe, have shown that the field of Language Resources and Technologies is mature enough to require consolidation of its foundations and assets.
The huge amount and diversity of language resources and tools, together with the availability of mature standards for content interoperability, suggests that the time is ripe for trying to weave the various resources scattered over different sites into a single organism of language services and repositories.
The integration and exploitation of language resources and tools into an architecture where users can combine elements of static language resources, such as lexicon, and dynamic processing resources, such as Natural Language Processing tools, is an active research topic pursued at several levels in the language resource interoperability field.
Nowadays, language resources and technologies, thanks to recent initiatives designed for making Language Resources available to specific communities, are more widely available than they were ten/fifteen years ago, but the entire Language Resources and Technologies (LRTs) community feels that two foundational building blocks for the future of the field are still either missing or they are in a very embryonic phase: we refer to the easy and fast access to information about LRTs and to the lack of well established standards to guarantee the interoperability among language resources and linguistic processing tools.
Recent initiatives in the LRTs community have been proposed to address these issues: CLARIN and more recently METANET, PANACEA are strongly based

on the construction, integration and maintenance of language resource catalogs as well as on the effort of defining standards for the production, processing, use and re-use of linguistic data.
Since the community of computational linguists is very wide-ranging, starting from human and social sciences scientists up to computational linguists involved in technical environments such as Machine Translation, it is aware of the impossibility of creating "*the Language Resource Infrastructure*" but that many Language Resource Infrastructures can be designed for solving different requests within the language resource field. Proposed infrastructures will have many common aspects, but every single infrastructure will have its own distinguishing characteristics, since it will be designed to solve specific problems.
In this thesis, we start describing key concepts typical of the Language Resource and Technology community and then we report the efforts carried out at our institute, the Istituto di Linguistica Computazionale (ILC), for both internal purposes and European projects, including the cited METANET and PANACEA to which our contribution are currently dedicated.
In particular, we will focus on the Authentication and Authorization Infrastructure and we will see how the *apparently simple* fact of "performing a registration" on an infrastructure will include a deep re-thinking of well-known concepts such as Identity Management, Access Management and Resource Management and will consequently constraint the actual architecture and modules of the infrastructure. Again, the (theoretic) building block of Authentication and Authorization Infrastructure will have more than one physical realizations.

Chapter 2

Language Resources: a brief introduction

In the last decades, Language Resources have became fundamental actors in Human Language Technology "also in view of developing innovative and robust technologies or to integrate existing ones to achieve more advanced applications". During the same period[1], LRs started to be considered as the platform on which new applications could be based: they started to play an infrastructural role as recognized by many European Language Resource (LR) projects during '90, [1]. Language Resources, to play an infrastructural role, need to be "reusable". This means that "quantitatively large"[2] LRs have to be designed according to various factors. Among these factors, we can cite the utilization of existing repositories (or catalogs) of information about LRs as sources for the construction of Natural Language Processing (NLP) systems, the construction of large resources designed to be used in different research areas and the design of "standards" to represent LRs, [2].

[1] We are talking of '80 and '90 of XX Century.

[2] The larger (in the sense of -strictly speaking- *size* is the resource, the more reusable should be, so that it can be used in different scenarios w/o being re-designed

2.1 Language Resources

We'll (ab)use the terms LR and Language Technology within this thesis. So it is time to give a clear definition of what these entities are or are intended to be. The most recent and accurate definition of a LR can be found in the European project FLaReNet. According to the view adopted and exposed in the project, the term LR "... refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving or evaluating natural language, speech or multimodal algorithms or systems ...", see chapter 3 and section 3.2 for more details.
The above definition of LR is applied to written, spoken, multimodal corpora as well as to lexicons and grammars, i.e. to LRs which mainly consist of data. The term LR, however, is now expanded to include basic software tools for creating, preparing, annotating, managing and using LRs (data). This kind of Language Resources (the ones that include software) are also known as Language Technologies. Both terms (Language Resource and Language Technology) are usually collected and shortened in the term Language Resource and Technology (LRT). Often, the more generic term LR is used for identifying both language resources and language technologies[3].
Over the past two decades, the Human Language Technology (HLT) community has recognized that language resources are one of the pillars of both HLT and HLT systems and that they have been strongly involved in the creation of computational lexicons, language corpora along with different linguistic annotation levels, and compendia of semantic information[4]

2.1.1 Language Resources and NLP

These three types of LRs, corpora, lexicons and semantic compendia, along with tools to manage them, represent the "core" resources for current NLP research. Current NLP research includes recent scientific developments in both the application fields of content management[5] and in the Human-Machine, Human-Human and Machine-Machine communication. These sectors are pretty active, nowadays, along with the corresponding theoretical areas, linguistics, cognitive science,

[3]Even in this thesis, we'll use the term LR for identifying the two kinds of resources, unless otherwise specified.
[4]Wordnets, Framenets, ontologies are examples of such compendia.
[5]Content Management entails different tasks such as, for example, content processing, access, creation ...

Artificial Intelligence (AI), robotics ....
This situation forces to broaden the definition of LRs, i.e. to re-define the coverage of the term to ensure a long-lasting credibility in a dynamic environment such as the one covered by recent NLP research.

2.2 Language Resource catalogs

The idea behind catalogs is to show how the Human Language Technology (HLT) domain structured itself under the incentives of "data centers" that initially collected information about LRTs and then started to catalog them. As a side effect, catalogs show how things in the LRT community evolved over the last decades, [3].
The second FLaReNet blueprint, [4], dedicates an entire section of its recommendations to the need for documentation of LRs according to common best practices: "... documentation is what makes language resources usable by people which did not create them..." In addition, the documentation of LRs should include information about data content and format as well as information regarding the context of production and intended applications and uses. The first version of the same blueprint, [5], affirms that: " ... documentation of language resources is generally poor. It is very difficult to find information about possible industrial uses of language resources ..."
Catalogs of LRs are information repositories which gather information on linguistic phenomena in different languages and domains. The mission of the various catalogs is to document as many LRTs as possible, so that all gathered information is not lost, and to ensure that documented LRTs will not disappear. The LR community, in fact, is aware that language resources that are not documented through a suitable set of key-words[6] do not exist or, at least, they are "invisible" to the community.
In following sections we briefly describe the major "data centers" both in and outside Europe which have carried out, during past years, the effort of documenting LRs along with a list of features useful for both the LR community and the industrial companies[7]. More information is available in [6].

[6] In following sections, we will see that these key-words are usually called *metadata* and play a crucial role in Language Resource Infrastructures.
[7] Additional information is available in the catalog official websites.

2.2.1 The ELRA Catalog and UC

The missions of Evaluation and Language Resource Agency (ELRA)[8] are "promoting LRs for the HLT sector, and evaluating language engineering technologies ...". Through ELDA[9], the official distribution agency, ELRA makes available all LRs described in its catalog. The ELRA catalog[10] includes all resources described according to a specific set of elements (metadata) which is strongly focused on the HLT community.
ELRA provides the Universal Catalog (UC)[11] as well; this catalog collects information about LRs identified all over the world. Its collection is not limited to LRs which are distributed through the Evaluation and Language Distribution Agency (ELDA) agency, but includes information on all Language Resources regardless their channel of distribution. The Universal Catalog is a repository for existing language resources along with their features such as legal issues, availability, intended uses, modality .... Changes to this catalog can be made by the ELRA team as well as by interested LRs producers.
The UC is constantly updated with all information related to its identified resources, so that the LR community is always informed with the last feeds. It is public since the 1^{st} October 2008, and both LR and HLT communities can freely have access to the information of existing resources (and to the actual resources if available in the catalog) in the world. So far, the UC contains more than 1500 LRs.

2.2.2 The Linguistic Data Consortium Catalog

The Linguistic Data Consortium (LDC)[12] is "an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes." Its catalog[13] is an American initiative correlated to LDC activities and includes, up to October 2009, $\sim$ 450 corpora of language data, including text, speech, video and lexicon resources, which are distributed through the LDC. The information about Language Resources is supplied by LRs

[8] http://www.elra.info
[9] ELDA is the acronym of Evaluation and Language Distribution Agency
[10] http://www.elra.info/Catalogue.html
[11] http://catalog.elra.info
[12] http://www.ldc.upenn.edu/
[13] http://www.ldc.upenn.edu/Catalog/

producers according to a set of elements and recommendations provided by LDC.

2.2.3 The Japanese NICT Universal Catalog

The National Institute of Information and Communications Technology (NICT) has been established by the Japanese government "... to carry out research and development in the field of information and communications technology, which supports the upcoming ubiquitous network society in an integrated manner from basis to application and also provides comprehensive assistance to the public and private organizations working in this field ..."[14].
The information regarding language resources collected by the NICT catalog[15] consists mostly in harvested data ("a la *OLAC*") from ELRA, LDC, GSK and other catalogs; NICT catalog contains $\sim$ 2700 LRs.

2.2.4 The LRE Map

The Language Resources and Evaluation (LRE) Map[16] is a totally new initiative promoted by the FLaReNet project, see section 3.2, and initially developed in collaboration with ELRA in conjunction with the Language Resources and Evaluation Conference (LREC) 2010. It was conceived as a campaign for collecting information about the Language Resource and Technologies underlying the scientific work presented at that conference. To collect this information, authors who intended to submit a paper were requested to provide information about the language resources either developed or used. The required information was pretty simple and related to basic information about the type of the resource, the language and modality represented, the intended or real application purposes, the degree of availability for further use, the maturity status, the size, type of license and availability of documentation, [7].
The new aspect of the LRE Map is, then, that information about LRTs is collected *by* the LRT community *for* the LRT community, according to a bottom-up strategy, so that this information represents the actual feelings of the community about the language resources.
The LRE Map soon became a very popular initiative joined by Conference on

[14] http://www.nict.go.jp/about/message-e.html
[15] http://facet.shachi.org/?ln=en
[16] A pilot interface on the LRE Map is available at http://www.resourcebook.eu

Computational Linguistics (COLING) and Empirical Methods on Natural Language Processing (EMNLP) conferences[17] and contains more than 2000 descriptions of resources. This shows that the idea has a great potential, and that the Map will easily become a a powerful "aggregator" of information related to language resources. So far, the LRE Map is a collection of metadata (see section 2.3) about Language Resources and Technologies collected in three major conferences[18] during 2010, but it is much more than a standard LRT catalog. In fact, in addition to information on language resources, the LRE Map gathers details about authors and papers submitted to the conferences, so that it is ready to become a social platform in the LR community.

2.2.5 The Clarin Virtual Language Observatory

The Virtual Language Observatory (VLO)[19] is an initiative started within the Common Language Resource Infrastructure Network (CLARIN) project, see section 3.3. From this service, interested users can explore the world of language resources and technologies cataloged in the the CLARIN inventory of LRTs[20] from different perspectives. The catalog can be browsed following traditional approaches, such as the original menu-driven viewing, and more advanced techniques based on both geographical and faceted facilities.

2.2.6 DFKI NLSR and LT-World

Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)[21] is one of the largest non-profit contract research organizations in the field of software technology based on Artificial Intelligence methods. DFKI is focusing on the complete cycle of innovation from world-class basic research and technology development through leading-edge demonstrators and prototypes to product functions and commercialization.

[17] Other conferences, such as International Speech Communication Association (Interspeech), Association for Computational Linguistics: Human Language Technologies (ACL-HLT), International Joint Conference on Natural Language Processing (IJCNLP) and European Association for Machine Translation (EAMT) have already agreed to use the LRE Map for their next year conferences (2011)

[18] LREC, COLING and EMNLP.

[19] http://www.clarin.eu/vlo/

[20] The CLARIN inventory is survey of LRTs, whose results can be found at: http://www.clarin.eu/view_resources for language resources and http://www.clarin.eu/view_tools for tools

[21] http://www.dfki.de, founded in 1998

DFKI provides Natural Language Software Registry (NLSR)[22] which is "... a concise summary of the capabilities and sources of a large amount of Natural Language Processing software available to the NLP community.". The NLSR is mainly focused on language technologies but catalogs language resourecs as well. In addition, DFKI provides the LT-World[23], a comprehensive portal intended to provide constantly updated information on LRTs.

2.2.7 The Japanese Gengo-Shigen-Kyokai

Gengo-Shigen-Kyokai (GSK) has been established in June 2003. Its main goal is to promote the distribution of both LRs and Technologies for contributing to Natural Language Processing (NLP) technology in both research and industrial development.

2.3 Metadata

Metadata can be defined as "data beyond data", since they provide a fundamental group of information which can be used to describe and catalog "objects". These objects can be physical, such as books cataloged in physical libraries, or digital objects, such e-book, documents, video, images, ..., cataloged in digital libraries. The history of metadata used to describe digital objects starts with the Electronic Text Encoding Interchange (TEI), [8], which developed a "standard for the representation of texts in digital form" and is culminating, in the last years, into initiatives which introduce the concept of metadata to cover a wider range of "objects". In fact, initiatives such as the ISLE Meta Data Initiative[24], Open Language Archives Community as well as the Language Resource catalogs described above (see sections from 2.2.1 to 2.4) are moving away from simple metadata schemas (Dublin Core, for instance) and proposing well defined linguistic concepts along with modeling constraints and standards like Moving Picture Experts Group (MPEG)[25] to manage multimedia content and define the complex landscape of the metadata as they are used nowadays.
The notion of metadata, as introduced by librarians, is now changing to include the semantics of described objects in order to manage the increasing amount of

[22] http://registry.dfki.de/
[23] http://www.lt-world.org/
[24] http://www.mpi.nl/IMDI/
[25] Version 7 available at http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

objects and the complex relations between them. However, the term metadata is ambiguous, since its meaning strongly depends on the context where it is used. In fact, metadata can be used to describe the data collected as well as to describe the objects we want to collect. For example, the same metadata, say, for instance, the *language* of a language resource, is used to collect all possible languages of a given resource (see above, section 2.2.4) and, according to this, the label *language* informs that the data collected under this label are languages, but the same metadata is used as a "dimension" (aspect) of the "object" language resource we have designed. In other words, the same single metadata can be considered as "structural" when it is used to design the container of data and "descriptive", when it is used to describe the single instance of a cataloged "object". In the latter sense, metadata are true data and can be used "to assist in using or interpreting other data ...", while in the former they are used to provide information about the design and specification of data structures we are collecting, [9].

We have seen in the sections dedicated to different language resource catalogs, that the common aspect of these initiatives is to gather information about collected Language Resources and Technologies. All this information is grouped into a set of categories which contain a list of coherent data. For example, *language*, *use* and *availability* (among others) are different categories which are used to identify possible aspects of cataloged LRTs and represent their metadata.

The manifold uses of metadata in the LR community prove that there cannot be one single schema which covers the requirements of all researchers, Language Resource and Technology producers, users, applications or systems. This aspect forces each LR catalog to define its own set of metadata, focused on the type of resources described: catalogs mostly dedicated to describe speech resources will focus on speech-specific metadata, such as the *sample frequency*, the *size*, the *format*, the *actor* ..., while catalogs which manage written language resources will use different metadata such as the *type*, the *availability*, *number of words* and so on, [10]. As a consequence, it is impossible to have a unique process able to collect metadata from different repositories and capable of merging different metadata schemes[26]. The most prominent gap that the LR community has to narrow is the abstract description of the basic LR categories along with their distinctive features and relations. To achieve this goal, metadata can be classified into metadata which constrain the object and control integrity and metadata purely descriptive

[26]This process is known as *metadata harvesting*. For more details, please visit http://www.openarchives.org/OAI/openarchivesprotocol.html.

which help to interpret the object, see figure 2.1. Further steps in the develop-

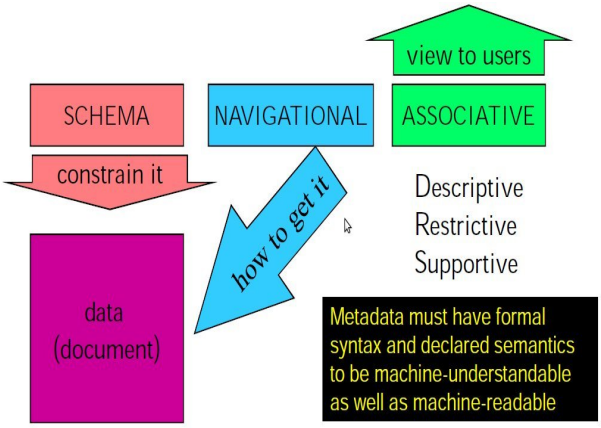

Figure 2.1: Metadata internal relations

ment of metadata will look at the definition of new schemes to describe software and services provided. In fact, new strategies provide a subset of metadata (the restrictive metadata in figure 2.1) especially designed to manage the use of the object e.g. rights, payment, uses, restrictions... The management of these services along with their access, can be enforced through such specific metadata, [11]. The following section lists down different initiatives related to metadata. Interested readers can get more information from [6].

2.3.1 Metadata Initiative

Data Category Registry The ISO Data Category Registry[27] is "an attempt to achieve interoperability among the various metadata schemas". Briefly, the interoperability is achieved through the registration of widely used concepts ("data categories") to guarantee a common terminology;

TEI The Electronic Text Encoding Interchange[28] is a consortium which developed a "standard for the representation of texts in digital form". Currently TEI is the most used one in the area of humanities;

[27] http://www.isocat.org/
[28] http://www.tei-c.org/index.xml

Corpus Encoding Initiative The Corpus Encoding Initiative (CES) applies the TEI philosophy to describe linguistic corpora. There is an XML version[29];

ISLE Meta Data Initiative ISLE Meta Data Initiative (IMDI) is a set of metadata used to describe specific LRs such as the multimodal. IMDI provides tools as well;

Dublin Core Meta Data Initiative Dublin Core Meta Data Initiative (DCMI) Dublin Core Metadata Initiative (DCMI)[30] is "an open organization engaged in the development of interoperable metadata standards that support a broad range of purposes and business models." The Dublin Core (DC) metadata set is a basic collection of 15 elements. The DC set is widely used and is a *de facto* best-practice to exchange metadata descriptions between various schemes: many metadata schema should have a DC core set, or, at least, should be DC-compliant in order to achieve interoperability.

2.4 OLAC

Open Language Archives Community (OLAC)[31] is "an international partnership of institutions and individuals who are creating a worldwide virtual library of Language Resources ..." The standard metadata set of OLAC uses the complete metadata set of DC (see section 2.3.1) to describe LRs[32]; the OLAC terms can be extended with extensions which are OLAC-specific[33] to better describe LRs. OLAC archives are harvestable, see section 2.4.1 using the OAI-PMH protocol.

2.4.1 Metadata Harvesting

Harvesting metadata means crawling metadata. Metadata harvesting allows for discovering and sharing resources across many repositories, [12].
Metadata harvesting uses the Open Archive Initiative Protocol for Metadata Harvesting[34]

[29] XCES, http://www.xces.org/
[30] http://dublincore.org/
[31] http://www.language-archives.org
[32] OLAC archives contain approximately 35000 records, covering resources in many languages (up to January 2011)
[33] OLAC provides the possibility to define further extensions.
[34] http://www.openarchives.org/OAI/openarchivesprotocol.html

protocol to share metadata between sites. This protocol has been designed to overcome the interoperability barrier and to unify incompatible and diverse metadata descriptions which can be found different repositories. The final goal of metadata harvesting is to create a large, virtual global repository of language resources and technologies.

2.5 Linguistic Annotation Process

Language Resources and Technologies are building blocks for developing NLP systems, [13]. These components, within a an NLP system, can be annotated at different levels for collecting information on linguistic aspects and for extracting new information to integrate in already existing resources, so that the newly created LRs are more complete than the old ones.
We focus here on a specific field the Natural Language Processing (NLP) task, the task used to add linguistic information to texts, since it is the main NLP task used in both METANET and PANACEA In this task, LRTs are used to add linguistic annotation[35] to a raw[36] text.
Essentially, the linguistic annotation process is a process where a raw text, for example a sentence, is analyzed by a software (for instance a POSTagger, a dependency parser ...) which reads the input raw text and adds linguistic annotations to produce an "annotated" text as output.
The process can be iterated so that new annotations are added. Figure 2.2 below displays a standard linguistic annotation process. As an example of linguistic annotation added to a row text, we can see the output of the *Freeling* tool[37] when a sentence ("el gato come pescados y mariscos") is analyzed, see table 2.1. From the output of the tool we can see that two linguistic annotations have been added: the lemma ("comer" is the lemma of the word "come") and the part of speech (the part of speech VMIP3S0 indicates that the word "come" is the (M)ain (V)erb whose form is the (I)ndicative (P)resent and whose person is the (3)third (S)ingular). According to figure 2.2, the process can be iterated to add more linguistic information. Freeling, for example, can add new linguistic information, since it have

[35] Briefly, a linguistic annotation is an additional (linguistic) information added to a piece of text.
[36] Raw text is pure text, without any formatting information such as paragraphs, comments ...
[37] Freeling is an open source suite of language analyzers developed at TALP Research Center, in Universitat Politècnica de Catalunya. http://nlp.lsi.upc.edu/freeling/

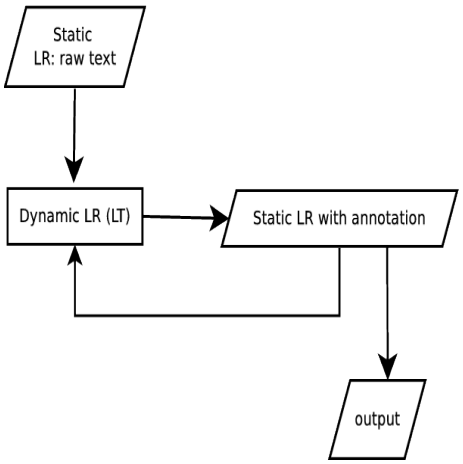

Figure 2.2: Linguistic annotation process: a raw text (a static LR) is (cyclically) analyzed by a tool (a LT) until a final output is produced.

```
el el DA0MS0
gato gato NCMS000
come comer VMIP3S0
pescados pescado NCMP000
y y CC
mariscos marisco NCMP000
```

Table 2.1: POSTagging of the sentence "el gato come pescados y mariscos"

been designed to perform various linguistic analysis on the same input text (in an iterative way): table 2.2 shows the output of the functional dependency of the previous sentence. Unfortunately, the linguistic annotation process is more complex than the one just showed in figure 2.2. In fact, usually, the linguistic annotation process is close to an assembly line, where different tools analyze both the input data and the previous annotations and add new annotation levels, as reported in figure 2.3.

```
grup-verb/top/(come comer VMIP3S0 -) [
  sn/subj/(gato gato NCMS000 -) [
    espec-ms/espec/(el el DA0MS0 -)
  ]
  coor-n/dobj/(y y CC -) [
    sn/co-n/(pescados pescado NCMP000 -)
    sn/co-n/(mariscos marisco NCMP000 -)
  ]
]
```

Table 2.2: Functional dependency analysis of the sentence "el gato come pescados y mariscos"

2.6 Language Resource Interoperability

The concept of "interoperability" will be discussed in following chapters, however a short definition can be provided here: Two language tools are interoperable if and only if the output of the first tool is one of the possible input of the second tool. According to this definition, the linguistic annotation process is strongly linked to the interoperability. During the process (for example the one sketched in figure 2.3) the second tool (LT2) can be invoked just after the first one (LT1) if the former (LT2) is able to read both the data and each annotation level added by the latter (LT1).
We will see that "interoperability" is a fundamental building block for Language Resource Infrastructures, as well.

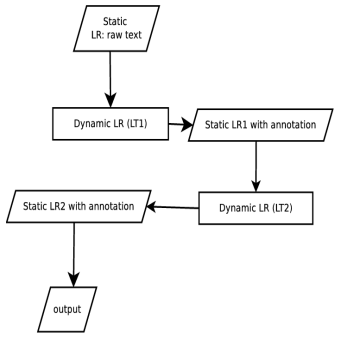

Figure 2.3: Linguistic annotation process: a raw text (a static LR) is analyzed by various tools (LT) until a final output is produced.

Chapter 3

Background

Different initiatives, starting from LIRICS and including recent FLaReNet, META-Net/META-SHARE and CLARIN have shown that the field of Language Resource and Technology (LRT) is mature enough to require consolidation of its foundations and assets.

Nowadays, language resources and technologies, thanks to recent initiatives designed for making LRs available to specific communities, are more widely available than they were ten/fifteen years ago, but the entire LRT community feels that two foundational building blocks for the future of the field are still either missing or they are in a very embryonic phase.

In fact, the community is conscious that the easy and fast access to information about LRTs in a readily and immediate way is a "must", but obtaining clear and verifiable information about language resources is still not easy to achieve, despite several catalogs are currently available (see section 2.2). In addition, the production, processing, use and re-use as well as the standardization of linguistic data represent a big effort for the daily work of Europe's industry and Small and Medium Enterprises (SMEs) connected to the community of LRTs, but, so far, there exists no established standard to guarantee the interoperability among language resources and linguistic processing tools, so that LRT experts are forced to adjust their data and tools according to different scenarios.

One of the concrete tasks that the Language Resource (LR) community has to face is to make these two foundational building blocks cooperate for creating "an open language *infrastructure* which allows networking of language technology pro-

fessionals and their clients..." [14]. In this infrastructure will be easy for any category of users [1] to find information about language resources in a fast, simple and immediate way from a main "entry point" This *infrastructure* will include both LTs and data sharing and it will based upon the *interoperability* key-word . In fact *interoperability* has been recognized as a very important pillar by different initiatives, and, nowadays, the LR community is aware that the forthcoming *infrastructure* can only succeed if the resources, tools and processes, which belong to the *infrastructure*, have been designed for working together in a seamlessly fashion.

3.1 LIRICS

Rationale

Linguistic Infrastructure for Interoperable Resources and Systems (LIRICS)[2] was a 3 years European project in the language resources and technologies field which recognized the importance of the use standards for assuring the interoperability among language resources and technologies during a linguistic annotation process (see section 2.5). Even if LIRICS addressed issues related to multilingual communication systems, the results achieved about the need for new standardization are, however, general and applicable to different domains in the LRT community.
One of the main differences between LIRICS and previous initiatives is that, in LIRICS, the standardization activities have been supplemented by open-source reference implementations (in different European languages) so that apply the standards to new languages and resources was easier. The LIRICS infrastructure has been designed to provide interoperability between existing resources based on the proposed ISO standards.

3.1.1 Project summary and description

LIRICS is one of the first European project in the language resources and technologies community whose main goal was to define a set of ISO standards for enabling interoperability and reuse of LRs, digital content and language engineering software (LTs). The project addressed the needs of information and communication society's scenarios, strongly based on multilingual communication, where the need

[1] LR community in first place, but also industrial players, funding agencies ...
[2] http://lirics.loria.fr; started in 2005

for new standardization as well as the recognition of existing *de facto* standards and their transformation into international standards was (and still is) increasing. The LIRICS project analyzed the available solutions used to facilitate the reuse of previously processed[3] language resources, [15]. LRs were annotated without any "standardized" tool; in other words there was no "common reference" to enable the exchange and reuse of data across different domains, languages and systems, i.e. the "interoperability" among LRTs was not guaranteed. However, the project recognized that the LRs enhancement and enrichment[4] were a *conditio sine qua non* for assuring the basic level of interoperability among LRs and that the standardization of linguistic annotations were the key solution for implementing it. Furthermore, for standards to really have impact, ordinary users[5] need to be able to both have easy access to the standards and to employ them without having to understand how they actually work.

Two of the main goals of LIRICS were, thus, to provide ISO ratified standards for language technology to enable the exchange and reuse of (multilingual) language resources and to facilitate the implementation of these standards for users by providing an implementation platform. Standards developed in LIRICS will lead to the optimization of the whole process of production, creation and sharing of language resources and will bring long-term benefits thanks to achieved interoperability and well documented tag sets. In addition these standards allow different coding conventions to be mapped to each other and to be compared across different corpora and different languages. Many current ISO standards started to be studied during the LIRICS project and have been provided to the European content and language industries, among them we can cite Lexical Markup Framework (LMF)[6], Syntactic Annotation Framework (SynAF) and Morphological Annotation Framework (MAF). These standards stimulate the reuse and standardization of terminology within and across reusable infrastructure that can be used in annotation projects, without need for further development, [16].

In addition, LIRICS defined a set of APIs to manage lexica through a common and standardized framework for the encoding of linguistic information to grant its reusability by different applications and in different tasks, [17, 18].

[3]By "previously processed" language resources, we mean (static) LRs that undertake a linguistic annotation process. See section 2.5 for more information and references.

[4]For example multi-level annotation for static language resources and new linguistic features for language tools.

[5]We mean users with no experience in linguistic.

[6]http://www.lexicalmarkupframework.org/

3.1.2 Contribution to the infrastructure idea

The contribution of LIRICS to the *idea* of Language Resource Infrastructure (LRI) has been twofold. On one hand, LIRICS defined the standards as the key feature for guaranteeing the interoperability among language resources through the organization of the managed data in fixed structures. On the other, LIRICS defined a set of APIs to manage these structures within an implementation platform. LIRICS started using web services as middleware between standards and the platform implementation. Web services implemented in LIRICS were developed to be consistent with both standards and APIs provided.

3.2 FLaReNet

Fostering Language Resources Network (FLaReNet)[7] is a European thematic network which aims at facilitating the interaction among the stakeholders of the field of language resources in order to drive a coherent evolution of the sector in the next years. FLaReNet intends to develop a common vision of such area to define a European strategy for consolidating the LRs sector and enhancing competitiveness at European level.

Its structure takes into account the various dimensions of LRs and the necessity of approaching them from different (technical, organizational, economic, legal, political, ...) perspectives. It also addresses multicultural and multilingual aspects, which are essential when facing the access and use of digital content in today's Europe. FLaReNet will consolidate the existing knowledge and contribute to structure the field of LRs of the future by discussing new strategies; its outcomes will help the LR community to identify the language resources of major interest, while blueprints of actions will be provided to the community as incentive - at both European and national level - to identify and develop new language policies supporting linguistic diversity in Europe and strengthening the market of language resources through new products and innovative services especially for less technologically advanced languages.

Previous experiences proved that networking is one of the privileged means to pool together major experts from different areas, reach consensus, make the community aware of the results and disseminate them in a fine-grained, pervasive way and this can only be achieved through a coordinated, community-wide effort to

[7] http://www.flarenet.eu; started in 2007

ensure contribution from the main actors of the various areas.
Again, FLaReNet addresses the challenges for digital content to become effectively usable in view of an inclusive information society: the development and exploitation of LRTs and their exhaustive documentation.

3.2.1 Contribution to the infrastructure idea

People in FLaReNet recognized the importance of documenting the language resources and technologies. The contribute of FLaReNet to the *idea* of LRI is that catalogs or, at least the easy access to catalogs, must be cornerstones for the infrastructure. FLaReNet indicates that legal aspects needed to manage non-free resources should be part of the infrastructure as well. FLaReNet does not suggest a technical solution for merging access to catalog, resource's availability and legal aspects[8]

3.3 CLARIN

Common Language Resource Infrastructure Network (CLARIN)[9] is a big challenging infrastructural project whose preparatory phase started in late 2008.
The idea behind CLARIN is that the LRT community is not ready for eScience since it lacks the pillars for a typical "research infrastructure". CLARIN is an attempt to change this situation since it aims at building and making operative a eScience infrastructure for the LRT community.
Part of the *mission* of CLARIN is "...the construction and operation of a shared distributed infrastructure that aims at making language resources and technology available to the humanities and social sciences research communities at large...".
The project intends to exploit the possibilities of what language resources and technology con add to the humanities and social sciences communities by making an analysis of the state-of-the-art situation in the use of language technology in this field and by using typical humanities projects as case studies for developing a research infrastructure oriented to the humanities and social sciences needs. In fact, CLARIN " ...aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling *eHumanities*..." by stimulating, in the humanities and social sciences communities,

[8] Legal aspects are somehow connected to the Intellectual Property Rights (IPR) issues.
[9] http://www.clarin.eu started in 2008

the use of language resources and technology to improve their research.
CLARIN points at creating a research infrastructure based on several pillars, including technical aspects as well as political. The technical objective, on one hand, is to provide a detailed specification of the infrastructure introducing concepts such as integration, interoperability, stability, persistency, accessibility, extendibility and a set of procedures to be adopted to make all this up and running on a validated prototype based on these specifications. The political objective, on the other, is to bring together the funding agencies and to establish an agreement between the funding agencies in the participating countries about governance, financing, construction and operation of the infrastructure.
The CLARIN infrastructure is based on the fact that language processing systems executed by computers are already part of many sub-disciplines in the humanities and social sciences fields. However, the cost of collecting and annotating large text or speech corpora, dictionaries or language descriptions and to digitalize them so that they can be managed by computers is huge and requires an effort that no single researcher in the humanities and social sciences can endeavor. Researchers can gain the benefits of computer-enhanced language processing only when there is a coordinated effort in creating a federation of existing archives and repositories of resources, and an infrastructure designed to provide access these resources along with the necessary tools to manage them.
The purpose of the infrastructure defined in CLARIN, is "to offer persistent services that are secure and provide easy access to language processing resources".

3.3.1 Contribution to the infrastructure idea

The big contribution of CLARIN to the *idea* of LRI is visible in the following aspects:

The concept of research infrastructure CLARIN allows researchers and ordinary users in Humanities and Social Sciences (HSS) to be part of a the eScience paradigm. This means that this kind of researchers will use systems and technologies usually utilized by other variety of scientists such as physicists, biologists, chemists ... CLARIN will prepare for HSS researchers an infrastructure based on secure grid technologies;

The concept of federation in a research infrastructure CLARIN has moved the concept of federation from standard information management, [19] to a

complex infrastructure. The idea of federating research centers in different countries by creating a pan-European super-structure has been very important for stimulating national initiatives for defining national federation such as, for example, the Italian IDEntity Management (IDEM)[10] Garr Federation;

The use of Single Sign On Single Sign On (SSO) is a cardinal rule in federation, since it allows users to log only once in the federation. The user identifier is sent from the CLARIN machinery to and from different centers in the federation;

The use of Persistency CLARIN hugely uses the concept of the Persistent IDentifier (PID) in its infrastructure. PIDs are generated by systems which have been designed to provide and maintain these PIDs for any type of resource, including software, web sites and so on. The idea behind PIDs is the possibility of cite the resource with this identifier instead of the name of the official web site of the resource itself. In this way, the resource is cited in a persistent manner with respect to possible variations of its name, location ..., [20];

Resource interoperability CLARIN accepts many ISO standards to describe language resources and technologies and, thus, to ensure the interoperability among them; in addition CLARIN formalizes a new level for guaranteeing such interoperability. In fact, CLARIN uses the ISOCat, see section 2.3.1 initiative for adding a sort of *standardization* of the (possible) values that the ISO standards can use to describe the language resources, [21].

In conclusion, in CLARIN users can access the infrastructure through distributed knowledge centers, [22, 23] and, using the single sign-on technology, they have the access to repositories of data with standardized descriptions, language tools capable of working on standardized data. All this scenario will be available on the Internet using a service oriented architecture.

3.4 PANACEA

Machine Translation (MT) is a strategic challenge to overcome language barriers while machine translation systems are expected to have a significant impact

[10] https://www.idem.garr.it/

for managing multilingualism. The PANACEA[11] project is expressly designed to address MT issues in Europe by making it possible to "...to translate the huge quantity of (written or oral) data produced, and thus, covering the needs of hundreds of millions of citizens ...".
The Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies (PANACEA) project is addressing the most critical aspect for MT: the so-called "language-resource bottleneck". MT technologies consist of tools which are generally language independent, but they depend on the availability of language resources which *are* language-dependent to be really applied in actual applications. The main issue for MT is then to supply MT system with "every pair of European languages, for every domain, and for every text genre". To provide such pairs of European languages, suitable LRs must be found, processed and supplied to MT systems developers. These LRs need to be provided in the format and with the information required by MT systems. In addition, since language is changing during time and there is often evidence of new knowledge in certain domains, language resources cannot be considered completed and what the MT community really needs is an automatic, and adaptive system for producing and validating language resources useful for MT systems. This integrated machinery for the production of LRs is exactly what PANACEA aims to build.
The objective of PANACEA is thus "to build a factory of language resources that progressively automates the stages involved in the acquisition, production, updating and maintenance of language resources required by MT systems ...". This automation of the process will reduce the time and the effort for creating language resources which are demanded by MT systems and technologies. In order to address these issues, PANACEA analyzes how to create a platform designed for managing dedicated workflows, created by the composition of a number of different web services especially designed for processing LRs and automatically produce a massive amounts of LRs required by MT systems and technologies.

3.4.1 Contribution to the infrastructure idea

The main contribution of PANACEA to the *idea* of LRI in visible on the formalization of web service workflow composition. PANACEA considers the web services as atomic services that are globally available and easy to be composed thanks to

[11] http://panacea-lr.eu/ started 2010

an interchange format defined in the project and used to model both input and output formats in a standard fashion. Web service composition is clearly based on the concept of interoperability and the PANACEA Traveling Object (TO) has been especially designed to make web services interoperable, [24].
PANACEA suggests the Taverna[12] Management System tool for web service workflow composition, however the guidelines that the project provides to the community regarding the TO can be applied to any other composition systems.

3.5 METANET

NETwork for the Multilingual Europe Technology Alliance (METANET)[13] is an ambitious European project, started in 2010, especially requested by the European Commission to address the needs about obtaining Information and Communication Technologies (ICT) applications at affordable costs. These applications are strongly requested for enabling communication, collaboration and participation across language boundaries, supporting each language in the advanced functionalities of networked ICT, so that users can have equal access to the information and knowledge society despite of language differences.
For these applications to be ready, several fields of human Language Technology (LT), such as Machine Translation (MT), including both automatic and machine assisted human translation, Information Retrieval (IR) and content production and management (among others) must advance in usability and availability. The main goal of METANET is to build a "...single EU information space reflecting and supporting the cultural diversity of our continent as an adequate foundation for the multilingual European information and knowledge society ...". Because of the big complexity of managing many different languages, this challenge needs a big effort of researchers and language communities as well as of several industrial sectors related to the LR community.
METANET will try to address the above challenges by approaching problems in collaboration with researchers of other fields including machine learning, social computing, cognitive systems, ... In addition, the project will mobilize European LRT community encompassing researchers through networking of researchers, developers and language professionals.

[12] http://www.taverna.org.uk/
[13] http://www.meta-net.eu/ started in 2010

In conclusion, METANET will prepare the ground for a large scale concerted effort of national and international research programs which can be used and enriched by LR communities and commercial technology providers.

3.5.1 METASHARE

The METANET project aims at creating METASHARE, an open platform where language resources and technologies are shared and provided to the LRT community. METANET has an entire part of the project dedicated to the construction of a LRI which is called Open Resource Infrastructure (ORI).
The contribution of METANET to the LRIs starts from the positions which summarizes what is happened in the LRTs technologies in the last decades. In fact, METANET recognizes that methods currently used in language technology research and development rely on the deployment and the wide availability of appropriate resources and tools. Unlike ten/fifteen years ago, today the paradigm of LRT spans almost all areas of language technology, including speech areas, technologies for extracting information from unstructured content, machine translation technologies development .... But, despite the strong dependence of research and technology progress on language and language-related data and tools, the landscape of LRT community is unorganized and highly fragmented. Although many European project (see sections before) addressed issues such as availability, accessibility and visibility of resources and tools, re-use, complex systems and service architectures projected for managing these issues lack of a multilevel interoperability. In conclusion, the field of language resources and technologies today presents problems at all three types of interoperability[14]: organizational, semantic and technical.
The main goal of METASHARE is to create an infrastructure for the LRT domain. Following current trends in information technology, METASHARE will consolidate and make best use of what exists in terms of infrastructures, data, tools, technologies and expertise, existing and emerging standards, and will provide an infrastructure that will be open, integrated, secured, and interoperable.

Open METASHARE is designed to be ever-evolving and scalable. This means, for example, that the number of resources and services, including free and for-a-fee, that form the resource base of the infrastructure will be increasing;

[14]The three levels of interoperability have been fixed in the European Interoperability Framework, see http://ac.europa.eu/idabc/en/document/3473

Integrated METASHARE will consist of distributed networked repositories and data centers accessible;

Interoperable the resource base will be standards compliant to overcome formats and both terminological and semantic differences;

Secured METASHARE will manage legacy issues such as IPRs clearing, legal compliance and secured access to licensable resources.

3.5.2 The METASHARE model

The targeted resources and technologies of METASHARE will encompass language data[15], language-related data, (possibly) associated to other media or modalities[16], language processing and annotation tools and technologies, (web) services for using such tools and technologies and eventually a complex workflow system for combining interoperable web services.
The target user base of METASHARE includes all possible users of LRTs, including academic as well as industrial practitioners, language professionals ...; METASHARE will provide services to different communities such as academic institutions, research organizations, individual researchers, national governments and SME developers and professionals. According to a different point of view, METASHARE intends to turn into a useful infrastructure for providers of language resources and technologies, users of these resources as well as LT vendors and language professionals.
This profiling of METASHARE is a key aspect in the infrastructure building process, since the platform provided in METASHARE strongly depends on specific user requirements which may change during the life of the project. The infrastructure is thought to be as wide as possible since the beginning so that it can cover different requests coming from different communities. The purpose of METASHARE is to interconnect the field of LRTs by launching a broadly-based, multilateral, scalable infrastructural platform suitable for the needs of both LRT providers and consumers.
To achieve this goal, the METANET consortium will adopt the following items:

[15] See section 2.1 for the definition of language data and related aspects.
[16] The repository system adopted in METASHARE is cabable of managing data in different media: data but also video, images ...

- adopt a flexible approach for modeling a complex dynamic system such as the proposed LRI;
- involve from the beginning a very large cross-section of interested parties, especially LRT providers and consumers;
- analyze existing models and their modus operandi;
- ensure a simple governance mechanism for managing the legal and organizational issues.

METANET will release stratified versions of the infrastructure, within its life span and beyond it. Two are the dimensions of this stratification: type of resources and technologies managed and steps of integration. The last aspect is the most strategic one, in fact METASHARE will start by integrating relatively few centers[17] and gradually extend to gathering more centers. METANET will study all possible models for connecting the centers which are (and/or will be) part of METASHARE: tight and weak classes of connectivity as well as possible infrastructural network connections and node functions will be analyzed. In doing so, it will assess the role of existing and emerging repositories, access points as well as the range of services to be offered.

3.5.3 METASHARE related projects and initiatives

METANET is strictly connected with concurrent networking and infrastructural initiatives such as Fostering Language Resources Network (FLaReNet), Common Language Resource Infrastructure Network (CLARIN) and Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies (PANACEA) By its nature, FLaReNet is an important METANET ally and collaborator in strategic and technical issues, as well as for mobilizing and "engaging" the LRT community. On a different side with respect to FLaReNet, CLARIN represents an important source of information and know-how in technical issues of standardization, metadata and interoperability since CLARIN owns a $3 - years$ experience even if this experience has been taken in the target user base of Humanities and Social Sciences (HSS) researchers which is focused more on content research and which is slightly different form the target

[17] The initial centers are the partners of the METANET consortium.

of METANET focused on technology development. Finally, METANET looks at PANACEA as the provider of workflows in the field of Machine Translation (MT).

3.6 Language Grid

Language Grid (LG)[18] is a Japanese project developed by National Institute of Information and Communications Technology (NICT). It "is an online multilingual service platform which enables easy registration and sharing of language services such as online dictionaries, bilingual corpora, and machine translations", cfr. [25]. The philosophy of Language Grid project is briefly reported in [26] and can be summarized as follows:
The Language Grid is a "Service Grid" designed for sharing language services which connects language service providers and users using web service technologies. Users and providers which want to use and/or be part of the Language Grid project must sign am agreement. Once signed, the agreement allows participants to provide the Language Grid with their services as well as to use and combine available services to create new services suitable for their needs.
The driving idea of the project is that a language resource, even a static data resource, can be transformed into a Web-based service, and hence effectively utilized through well designed access interfaces. This trend would open up a new dimension for sharing language resources and technologies [27], which would also solve/remedy non-technical issues, such as intellectual property right. According to these aspects, Language Grid provides a "place" where LRT providers can define their resources as web services and make them available to a wider community. Language Grid aims at offering the following main benefits to the LRT community: facilities for both combining language resources and/or technologies and for adding own language resources to create new language services, [28].

3.6.1 LG architecture

Language Grid is an Internet-based infrastructure which allows a better intercommunication among people from different countries which share content in different languages. Its architecture is very complex, see figure 3.1, and merges horizontal with vertical elements, [29]. The bottom layer of the architecture, and the most innovative, is the *P2P* grid infrastructure. This layer is aimed at connecting LG

[18] http://langrid.nict.go.jp/en/index.html

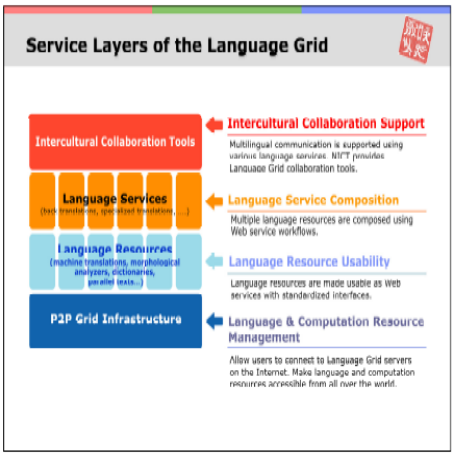

Figure 3.1: Service Layers of the Language Grid.

nodes and combining language resources on the web. Language Grid defines two types of nodes: core and service. The former manage all registered language services and provide features to search and compose simple language services. Web service composition is based on workflow which is managed my core nodes as well. The latter nodes are the "place" where language resources are deployed as web services.

The *P2P* layer is responsible for all (registered) information of language resources to be shared among all core nodes, so that the same services are equally available, regardless of which core node users access.

The "Language Resources Layer" is where language resources will be deployed as web services. Language Grid provides software APIs which help developers to make their web services consistent with the LG machinery; these APIs form standardized interfaces for given services such as morphological analyzers, dictionaries ... LRTs providers use these interfaces to release their services to the world.

The next architectural level is the "Language Services Layer". This layer is responsible for web service workflows. The last level, the "Intercultural Collaboration Tools", is designed for final users. This level contains different tools, including the Language Grid Toolbox[19] which provides a series of intercultural collaboration tools for supporting multilingual communities. It is developed in the framework

[19]http://langrid-tool.nict.go.jp/toolbox/

of "Service Grid Open Source Project[20]" and it is opens to everybody.

[20]http://servicegrid.net/oss-project/

Chapter 4

Language Resource Interoperability

We have seen in chapter 3 that the Language Resource (LR) interoperability played an important role in various old projects and it is a cornerstone for modern ones which establish the interoperability as the key for guaranteeing the sharing of language resources. We can "try" a definition of LR interoperability as follows:
by "resource interoperability" we mean that two (or more) resources can be combined in a workflow fashion.
Resource interoperability is usually defined at two distinct levels: a high level interoperability which addresses input/output issues, normally related to the *structure* of the exchanged information, and one low level interoperability which manages the *content*, i.e. the actual domain of the exchanged information. Usually, we can refer to the high level interoperability as to the *syntactic* interoperability, while the low level interoperability is called the *semantic* one, [30]. The low level interoperability is called *semantic* since it concerns the semantics (content) of the interchanged information. The use of the Data Category Registry catalog is fundamental to guarantee a *semantic* interoperability.
Pioneering works on resource interoperability started in the 90*s* with the Eagles[1] and ISLE[2] projects, whose results have been used in LIRICS and then consolidated into ISO standards.

[1] http://www.ilc.cnr.it/EAGLES96/home.html
[2] http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

4.1 Related approaches to LRs Interoperability

Language Resource (LR) interoperability can be managed from different perspectives. Among others, we can cite the Unstructured Information Management Architecture (UIMA)[3], [31], and General Architecture for Text Engineering (GATE)[4], [32], frameworks.

UIMA is software system that analyzes large volumes of unstructured information in order to discover knowledge that is relevant to an end user. It deals at language resources as software "hooks" that can be "handled" by a common framework; the UIMA platform provides facilities for embedding tools and resources into an Integrated Development Environment (IDE), such as Eclipse[5] and defines an object, the Common Analysis Structure (CAS), which contains both the input physical data (document(s)), metadata and any annotation level, (the *features structure*)[6], added by linguistic tools. It provides cooperating UIMA components with a common representation and mechanism for shared access to the document(s) being analyzed.

In a standard workflow, the CAS object runs from the input to the output steps and is accessed, manipulated and updated by each resource. This behavior of the CAS object allows developers to consider the UIMA workflow as an *assembly-line*. Developers, therefore, can choose the step as well as the conditions for a tool to be executed within the *assembly-line*. The UIMA framework manages resource integration by defining specific "descriptors", i.e. XML files. One of these files contains annotations performed on the content of the documents, while another contains the framework-provided infrastructure (primitive analysis engines) that allows them to be easily combined in a workflow (aggregate analysis engines).

GATE is an architecture and a framework for managing LEs. The *architectural* aspect of GATE is used to define how an Language Engineering (LE) system is organized, including how components interact each other and if these interactions satisfy the overall system requirements. As a *framework* it provides a reusable framework for managing LE software systems and a set of core libraries. In addition GATE provides reusable LE modules which are able to perform basic Natural

[3] http://uima.apache.org/
[4] http://gate.ac.uk/
[5] http://www.eclipse.org
[6] UIMA defines type system for document annotation. Briefly, a type system is a schema or a model for the CAS object. It defines the types of objects and their features (capabilities) that may be used by a CAS. Analysis engines conform to a type system.

Language Processing (NLP) operations. As UIMA, even GATE uses the standoff annotation technique to add information to documents analyzed and implements Annotation Graph (AG), [33], to manage the annotations.

4.2 Standards and Interoperability

Both UIMA and GATE analyze documents and add linguistic information as standoff annotations. As we have seen in section 2.5, two language tools are interoperable if and only if the output of the first tool is one of the possible input of the second tool, i.e. if the second is able to manage the structure (high level interoperability) and understand the content stored in the annotations (low level interoperability) of the document encoded by the first tool.
As consequence of this aspect, the first step toward a resource interoperability is carried out by defining a mature set of standards to be used for describing possible input and output formats. This is the area where ISO standards came to play. Standards such as LMF, MAF, SynAF, ... can be used for structuring the annotation schema(s) (again the high level interoperability).
One possible solution for managing input/output standardization is the introduction of a pivot standard, for both input and output formats. This mapping mirrors a graph structure in which each distinct format is mapped onto the pivot one, rather than onto every other possible format belonging to the same graph. This solution has been used in [34], where the Graph Annotation Framework (GrAF), [35] is used as a *lingua franca* to manage interoperability in both UIMA and GATE frameworks.

4.2.1 Language Resource Interoperability and Metadata

Current objectives in the metadata research field are committed to ensure interoperability and to explore compatibility issues as well as to remedy gaps in the LRs production and management in and for the Language Resource and Technology (LRT) community.
The lack of a formalized and abstract description of basic Language Resources, see section 2.3, has brought researchers involved in the CLARIN project (see section 3.3) to define a component-based metadata schema which is used to describe basic aspects of generic LRs in terms of reusable profiles. CLARIN Meta Data

Infrastructure defines profiles for lexicons, tools and web services as well. As reported in [36] " CLARIN Meta Data Infrastructure (CMDI) allows users to design their own set " of profiles, which are tailor-made and specific for their purposes, "as long as they make use of widely agreed concepts that are stored in the ISOcat registry and therefore guarantee interoperability". The interoperability is guaranteed at both syntactic and semantic levels. Syntactic interoperability is guaranteed by providing the structure of the profiles for language resources which must be consistent with other (ISO) initiatives such as ISLE Meta Data Initiative (IMDI) and Dublin Core (DC); semantic since the Data Category Registry (DCR) (ISOCat) includes all metadata concepts needed by the LR community. In practice, CMDI has moved from syntax to semantics in order to achieve high levels of interoperability.

Chapter 5

Language Resource Infrastructures

The creation of open and distributed "linguistic infrastructures" for Language Resources and Language Technologiess, is a new trend in recent RnD projects. These infrastructures are based on sharing (language) resources and tools as well as on a common knowledge for dealing with sharing-related aspects; thus, it is very urgent to create frameworks which can combine technological and organizational aspects to enable the cooperation among many groups which will work on the broad panorama of sharing resources: to merge results; to make them accessible to various applications; to empathize the use of standards for guaranteeing interoperability; to manage legal and Intellectual Property Rights (IPR) issues and so on, [1].

Starting from 6^{th} framework program, written, spoken and, currently, multimodal Language Resources are being playing the role of strategic components of "linguistic infrastructure". The availability of adequate Language Resources, see section 2.2, for "as many languages as possible is a pre-requisite for the development of a truly multilingual Information Society ... Language Resources have been recognized as a priority within a number of national projects around Europe", as reported in [37] during [38].

Nowadays, there is a huge amount and diversity of language resources and tools; this aspect, together with the availability of mature standards for content interoperability, suggests that the time is ripe for trying to weave the various resources

scattered over different sites into a single organism of language services and repositories.
The integration and exploitation of language resources and tools into an architecture where users can combine elements of data language resources, such as lexicon, and Natural Language Processing (NLP) tools, is an active research topic pursued at several levels in the language resource interoperability field.

The Istituto di Linguistica Computazionale (ILC)[1], is involved, both independently and in the framework of European and international context, in projects which address these subjects, see chapter 3.
In this chapter we start describing the Language Resource Infrastructure (LRI) planned for the ILC and, then, we will add our contributions to several projects, including METANET and PANACEA to which our efforts are currently dedicated.

5.1 The ILC Infrastructure

The architecture of the planned Language Resource Infrastructure is centralized with client server functionalities. We designed a central server which plays the role of a the central authority in a FDBS (see below) scenario. This server is internally accessed by the UIMA framework which provides services for users who want to access the infrastructure, see figure 5.1.
We approached the integration of Language Resource and Technology available at our institute using the Federate Database Architecture System (FDBS) technique, [19, 39]. This approach manages resource interoperability issues as well as resource structure definition and cataloging. FDBS defines a neat central authority responsible for overseeing all the interoperability outcomes among components and for defining an input/output standardization for resource communication protocol. This approach is preferred to a standard resource-sharing architecture, since the FDBS approach manages resource-sharing issues as well as users and roles definition. The central authority, defined by FDBS, oversees to the federation policies such as internal rules, groups and user rights, components cooperation and composition through a set of specialized registries used to address specific topics such as internal resource structure, resource-resource interaction and single resource role within a complex NLP workflow. The basic idea is that users can define their NLP

[1] http://www.ilc.cnr.it

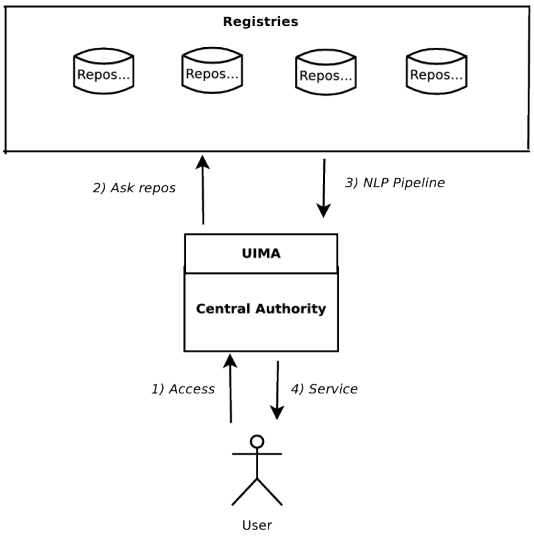

Figure 5.1: Simple scenario for the Istituto di Linguistica Computazionale Language Resource Infrastructure.

pipelines just assembling one resource after another. The central authority manages the mapping of internal formats onto the pivot one, while other repositories apply rules to compose NLP pipelines from single components.
We pointed out three main characteristics an architecture should have to be considered "federable":

Heterogeneity: heterogeneity in resources is due to several factors. Among them, differences in structures and semantics of data;

Autonomy: in a few words, autonomy stands for the quality, for a resource, to function independently from others;

Distribution: resources exist before the federation is built. The federation aims at the interactions of each single resource with others, defining interoperability rules, access rights and a common access language.

The infrastructure designed for the ILC addresses the issue of developing an interoperable infrastructure for language resources and technologies. We extended the FDBS adding typical functionalities coming from UIMA, see section 4.1 for details on Unstructured Information Management Architecture (UIMA) and section 5.2. In this way, we capitalized the advantages of a federated architecture, such as

autonomy, heterogeneity and distribution of components, monitored by a central authority responsible for checking both the integration of components and user rights on performing different tasks. We used the UIMA approach to manage and define one common front-end, enabling users and clients to query, retrieve and use language resources and technologies which are provided by proposed LRI.
With UIMA, we moved from the FDBS to a Federate Resource Architecture System (FRS), [40], simply defining a registry of available components (both static resources such as lexicons and corpora and dynamic ones such as tools and general purpose language technologies and allowing UIMA to play the role to the central authority, see section 5.2.1.

5.2 UIMA approach to FDBS

The administration of a Federate Resource Architecture System is a challenging task, since, in this architecture, the central authority has to manage both user rights and resource interoperability. User rights join interoperability rules and define a complex scenario in which the 'motto' "*who can do what and how*" is the key question.
We introduced specialized repositories to manage language resources[2], Natural Language Processing pipelines and user requirements defined in and for the federation. We have defined six different repositories and pulled apart resources from workflow and user information. Resource types as well as their capabilities are stored in differentiated repositories. By *capability* we mean a specific functionality of one given resource and, by generalization, the piece of information added to the input document by that capability.

meta-repository: this repository contains information on each single resource. Each resource is assigned a unique persistent identifier. This repository keeps track of each *resource instance*: a *resource instance* is a physical copy of the resource identified by the unique persistent identifier;

resource type repository: this repository classifies the resources in different resource types in agreement with resource capability;[3]

[2]Language resource survey is a prerequisite for the architecture to be made up.
[3]Resource types can further be sub-typed: for example, if a Tagger-type-A resource has three annotation levels, Lemma,Pos and Morphological information and a Tagger-type-B has only two

resource-schema: this schema describes which kind of data is managed by a language resource as well as input/output and standard compliance;

meta workflow schema: this schema represents possible workflows within the infrastructure. One single workflow schema assembles a list of resource types ordered according to execution priorities. The meta workflow schema provides a skeleton for NLP pipeline. The advantage of this schema is that we can define workflows using resource type as building block. Sometimes, a workflow can be specified upon peculiar resources;

user rights repository: this repository contains privileges assigned to users and/or groups. User privileges are defined both at resource type and resource level. User rights (or privileges) have to be placed within an *Identity Management* scenario and addressed with specific technologies [41, 42];

federal dictionary: this is a specialized component, which regulates the infrastructure topology. The federal dictionary is built upon other repositories described above and manages the resource taxonomy and interactions.
Every repository here described is defined accordingly with metadata, see section 2.3 directives; this set of repositories is the backbone of the architecture: it is used for both resource querying and services providing. Repositories defined above are internally accessed by the UIMA framework and, externally, by users who want to build their own resource collection relevant for their research. It is straightforward to identify one single language resource with a primitive analysis engine[4]. These analysis engines can be deployed as web services, since this is one of the deployment options supported by the UIMA framework.

Figure 5.2 describes the interactions defined among the repositories.

5.2.1 UIMA Role

UIMA is a framework designed to manage resource interoperability and integration in a corporate research environment; it is the obvious candidate to carry out

of the above levels, then the latter is a sub-type of the former. In other words, the resource type repository defines an ontology of language resources and technologies.

[4]Here, we refer to language tools, since UIMA analysis engine model more dynamic than static resources

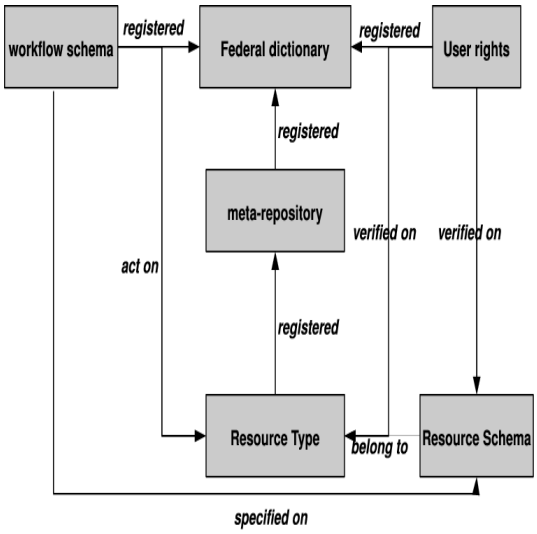

Figure 5.2: Interactions and relations at repository level.

the *role* of the *central authority* in a resource federation.
We used the UIMA framework to access and manipulate documents. In this basic scenario, the Common Analysis Structure (CAS) object is instantiated upon the document and it is accessed, manipulated and updated by analysis engines.
In order to define the role of UIMA within the ILC platform, we decided to add *operational annotations* to the whole document. In such a case, the CAS contains the document to be analyzed ***and*** the linguistic annotation a user needs, expressed as *operational annotations* defined at document level. Each *operational annotation* records linguistic annotations and/or language resources pertinent for these annotations and is an extension of native UIMA type systems, see section 4.1 and references for UIMA cited in that section.
UIMA played the role of the software interface between users and the ILC infrastructure; it is responsible for accessing the federal dictionary, selecting the right resources to perform linguistic annotations and return the results to the user. UIMA behaves just like a *central authority*: it tries to build an NLP pipeline according to user requests by checking if user requests are consistent with information registered into the repositories in terms of user rights, resource availability and so on.
Moreover, the choice of UIMA as *central authority* had helped in defining a stan-

dardization for input/output formats. In fact, the UIMA framework provides an out-of-the-box standardization of the structure of the analyzed document[5]. The standoff annotation can be accessed external tools or saved in a database as well, and provides a standard format to address data transfer from one resource to another. What we needed (and need already), at least from a linguistic perspective, is the *semantic* interoperability. We have used the Data Category Registry (DCR) [43] to partially solve this issue by defining a controlled vocabulary that limit the values of linguistic annotations.
Figures 5.3 and 5.4 describe the "annotation framework" we defined for the ILC Language Resource Infrastructure.

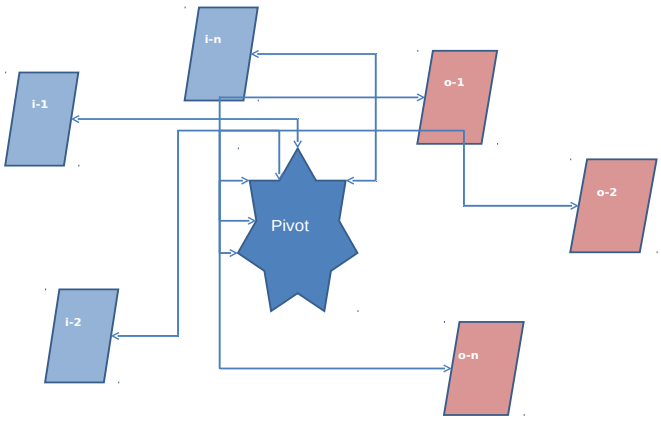

Figure 5.3: Graph structure for input/output mapping.

The "annotation framework" includes a a pivot standard for managing input and output formats and graph structure to map each format to the pivot one. This mapping scenario is accessed by the UIMA framework and managed by CASes.

[5]UIMA produces an XMI file which represents a standoff annotation, i.e. a type of annotation where linguistic information are put at the end of the file and linked to the documents through a set of identifiers. Standoff annotations ensures *syntactic* or high level interoperability, see chapter 3.

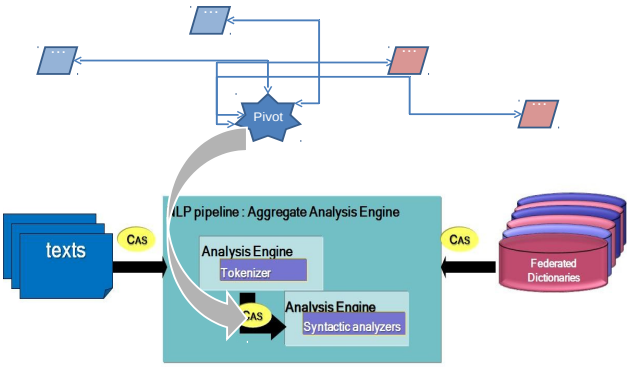

Figure 5.4: Graph structure is accessed by UIMA and managed by CASes.

5.3 LRIs: General considerations

There is not only one Language Resource Infrastructure but many Language Resource Infrastructures. In fact, LRIs are all alike, since they have many common aspects, but every single LRI is peculiar in its own way, since it has its own distinguishing characteristics.

The definition of Language Resource Infrastructure, thus, is not clear, even if some aspects are well-defined. We will try to define both common and specific aspects of LRIs in a *bottom-up* fashion, starting from what the Language Resource (LR) community feels about LRIs and then mapping their requirements on possible technical solutions. In fact, the Language Resource and Technology community probably knows perfectly what an LRI should offer to them but is unable to design it deeply in details technically; the "joint venture" between Language Resource and Technology (LRT) and Information and Communication Technologies (ICT) communities, which is common in recent projects, attempts to solve the dualism between linguistic *desiderata* and technical issues, testing actual technical solutions on typical linguistic scenario.

The most common way to gather the feelings of LR community is to perform interviews, [12].

The LR community consists of two main actors: *providers* and *consumers* of Language Resources, during the interviews, both actors of LR community are asked to provide their *desiderata* on LRIs. The interviews, however, do not distinguish between common and specific aspects of a Language Resource Infrastructure so that ICT experts must extract the general aspects from the big amount of data and, at the same time, keep track of requirements which can emerge from specific scenario. For example, the need of a registration module is a general aspect, but the actual implementation, including architecture, model, software to use, may depend from scenario to scenario.
From the interviews, we have found that *Consumers* of LRs usually ask for:

Registration Consumers do not prefer to register to the LRI, but can accept the registration upon some benefits such as personalization options, more resources available ...;

Search Consumers need to query for a LR using both keyword-based searches and through menus which allow to browse for provided LRs;

View Consumers will have a view on each LR available in the LRI. Consumers want to decide whether a LR is relevant from information provided;

Licenses Consumers think that if LRs are for-fee subscribe, they need to be advised of licensing issues and restrictions imposed on the use on a LR. Moreover, they do not be bored of licensing matters; they want the LRI manage it for them;

Obtain Consumers want to access the LRs through two modalities: direct download of the resource and via web services;

Feedback Consumers want to give different kinds of feedback[6] to the LRI. They believe feedback are useful for new consumers and providers as well;

Language Technology Consumers consider the exploitation of language tools through the LRI as a clue. LTs available in the LRI should overcome copyright issues.

On their side, *providers* of LRs express similar requests:

[6]Feedback will most probably regard quality, errors and solutions to problems.

Registration Providers believe registration to the LRI should be mandatory. They think registration could be twofold: personal (for a single provider) and institutional (when providers register LRs which belong to an academic or industrial institution);

Description Providers will provide a description of their LRs;

Upload & Publish Providers need to upload the LRs and make them available to the community. Providers require that this process is "fast". In addition, they think that the LRI needs to notify that (new) LRs are available;

Licenses Providers will choose the licensing terms for the distribution of their LRs. They expect the LRI can help them in choosing licensing and solving IPR issues;

Monitoring Providers are interested in monitoring their LRs. They are specifically interested in successes, failures, uses and numbers of downloads;

Feedback Providers will be able to get or access feedback about their LRs;

Evaluation Providers expect the LRI provide services for evaluating their LRs;

Update (Versioning) Providers may need to update their LRs. They are interested to a versioning system to keep track of old versions.

One different type of actors in Language Resource Infrastructures is represented by the physical sites, also known as data centers or repositories, where providers release their resources and consumers consume available LRs. These sites can be geographically distributes, such as in a European network of repositories, or centralized. Regardless of such distinctions, both providers and consumers assign data centers the following task:

Long Term Storage Data centers will provide long term preservation of stored resources. This aspect is connected to the versioning for LRs;

Language Resource Classification Both consumers and providers believe that language resource classification will help the LR in organizing itself in the future;

Promotion Data centers will widely promote stored LRs.

As conclusion, Language Resource Infrastructures are expected to support both providers and consumers. Providers expect support both in terms of exchange formats and in terms of laws for licensing; in addition, they expect that the LRI will take care of the distribution of LRs so that researchers[7] can concentrate completely on research and on the development of their Language Resources. Consumers wish to get from the LRIs an easy accessible, clear overview of the resources available; the more complete and accessible is the information about a resource, the easier is to understand what is relevant for the consumer (without the need to read a lot of documentation). In addition, consumers expect the LRIs will provide a good search mechanism. Finally, LRIs will collect various data centers under the "official" umbrella of the infrastructure to make clear to users about where and how to get resources from the different data center: users only have to join the LRI to access resources, regardless of where these language resource actually are stored.

[7] Researchers are currently the providers of their resources.

Chapter 6

Identity and Access Management in LRIs

Consumers, providers and data centers, need to register to Language Resource Infrastructures in order to use provided services, see section 5.3. These three actors will be the main groups of many LRIs users. Each group is interpreted as a different role and different roles have different set of rights, copyright restrictions as well as interface settings, profiles and so on. User registration is usually offered by data centers, which belong to LRIs, and it encompasses a set distinct activities as reported below:

Identity Management (IM) By IM, we mean the tasks required to manage a user's identity within a structured organization. Typical operations of IM are the creation, updating and cancellation of users. Identity Management is the place where human agents are defined into machine entities. In distributed environments, each single IM should be able to connect to other IMs systems;

Access Management (AM) By AM we mean tasks related to the user authorization process. These tasks provide different access levels to offered resources and services based on both rules (defined within the LRI) and users that actually request to access the resources. AM is responsible for checking whether the accessed resource is protected. In this case, AM requires that the requesting users have to be firstly authenticated on the IM system, so that their privileges (as provided by IM) will be checked with respect to the

ones needed to use the protected resource.
Many access management tools implement the "Access on Demand technique", for example [44]. This technique requires that privileges have to be checked when needed instead that at the "beginning of the transaction", i. e. when a user joins (logs on) the Language Resource Infrastructure;

Resource Management (RM) By Resource Management we mean the tasks needed to register, catalog, profile the resources which belong to the infrastructure;

Single Sign On The concept of Single Sign On entails that once a user logs into an infrastructure, (s)he retains all rights deriving from the respective user account. Attributes such as user rights, copyright restrictions, ... are passed from internal data center to internal data center. In other words, this means that the user will not be obliged to enter username and password each time (s)he decides to access a resource from data centers.

Identity and Access Management play fundamental roles in setting up Language Resource Infrastructures since the resource access and integration is strictly connected to how and by whom these resources are accessed.
Infrastructure designed to manage distributed language resources and to provide lexical services over the network to end users need robust identity and access paradigms to be implemented. In fact, as web technologies started to increase, the models for delivering information and services within the Language Resource Infrastructure (LRI), trough the network dramatically changed, requiring more complex Authentication and Authorization Infrastructure systems in a distributed environment. We have seen in chapter 3 how recent initiatives consider the provisioning of linguistic functions of lexical resources, as lexical services via the Web in a distribute environment as a prerequisite. The more distributed is the infrastructure the more linked it is to the management of identities which request to access a specific resource in the infrastructure, as well as to the many ways can be used to grant/deny those identities the access to available resources. Moreover, identity and access management. As we have seen in section 5.3, *providers* and *consumers* have pointed out the aspect to be released from managing privacy, security and legal issues when they join the LRI, requiring that the LRI should be the entity selected to manage them, allowing the users to access the LRI and automatically grant access only to the resources that they are enabled to use.

IM and AM need to be addressed before the Single Sign On, since the latter is in charge of passing user rights within the infrastructure, these user rights must be previously defined (which is what IM and AM aim to). Both IM and AM can be considered as two distinct processes, but the can be considered also a unique combined process (Combined Identity and Access Management (CIMAM)).

6.1 IM and AM as two distinct processes

Identity and access management, when implemented as two distinct processes, can be considered redundant. In fact, in many (even industrial) solutions that offer access to resources, the Access Control List and related groups are defined and managed by access management modules at resource level[1]. This implies that the same access groups which are managed in Access Management modules must be handled by the Identity Management module, as well, in order to be consistent. As a consequence, these solutions present a *twilight zone* in which groups defined in the AM system are assigned to identities (in the IM module) via the membership concept.

6.2 Combined Identity and Access Management

We have analyzed a Combined Identity and Access Management, in order to have a single "security checkpoint. This single "security checkpoint is part of the central authority (see section 5.1), and guarantees both flexibility and scalability. Such approach simplifies the management of the *twilight zone* since decouples the identity-related issues from the resource access problems. ACLs are defined once at IM and AM level, and, then, exported as metadata at the resource level.
The Combined Identity and Access Management can automate and simplify the management of user identities, access rights and compliance policies across the organization since it centralizes security management and makes it easy to deploy secure applications.
Our implementation of a combined identity and access management comes from the analysis of this simple workflow. In a basic scenario, the users join the LRI to find out which language resources can be used for specific NLPs tasks. If the

[1]This means the Resource Management must implement a module for managing ACLs. Usually this is implemented at matedata level (see section 2.3)

users find out that, available resources cannot be used for the specific purpose, then users should be able to combine such resources to create a new one suitable for their specific tasks. The workflow consists of the following basic steps but can be extended to a large amount of real situations:

1. The user joins the infrastructure;
2. The user obtains credentials and privileges;
3. The user accesses available resources;
4. Finally, user credentials and privileges are checked with respect to privileges and credentials needed for accessing each specific resource.

On the contrary to commercial implementation (which implement the "on demand" strategy), we decided to define the access credentials at the "beginning of the transaction", that is to say when the user join the LRI. This means the Combined Identity and Access Management provides the user all privileges along with resources that (s)he can access when (s)he joins the infrastructure. We have defined a list of groups and a set of Access Control Lists. As second step,these ACLs have been assigned to groups. Users, finally, belong to groups thanks to the membership attribute. When the user accesses the LRI, the identity-access management tool provides a list of ACLs according to the groups to which the joining user belongs. The incoming ACLs are then mapped on the Access Control Lists defined at metadata level for available resources. Figure 6.1 describes this solution. The solution we analyzed is flexible, since new users and groups can be automated and simplified by the use of a single administrative task across the whole platform. In addition, this solution simplifies the Resource Management tasks: a new resource in the LRI is a new record in the registries of the infrastructure and its access is ruled by the ACL2MD module in figure 6.1.

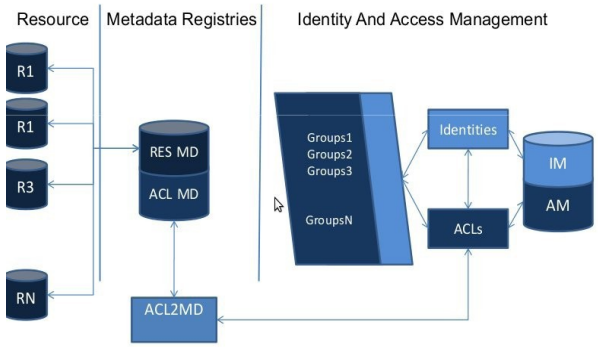

Figure 6.1: Combined Identity and Access Management.

Chapter 7

SSO in Language Resource Infrastructures

7.1 Single Sign On

Usually, Language Resource Infrastructures are structured and geographically distributed[1] infrastructures. As a trusted network of repositories, the LRIs require a specific management of logged users, called Single Sign On[2]. In such infrastructures, users should be able to login:

- by using one single identity;
- by logging in only once;
- by signing network-wide service provider conditions only once.

Security Access Markup Language (SAML)[3], developed by the OASIS Security Services Technical Committee, is an XML standard for exchanging authentication and authorization data between security domains. It provides standard mechanisms for organizations, called Identity Providers (IdPs), which are used to authenticate users on the infrastructure and Service Providers (SPs) used to provide

[1] At least the LRIs which are the outcomes requested by the European Commission in recent projects.

[2] See appendix A for details on a specific implementation of the Single Sign On (SSO), the one provided by *Shibboleth*.

[3] http://saml.xml.org/

services. According the SAML, it moves "...packets with attributes of people rather than data" to and from IdPs and SP.

7.2 METASHARE Architecture

Before discussing the SSO in METASHARE, we have to briefly present the architecture that is currently under development in the project, as reported in [12]. The architecture so far proposed for the METASHARE platform is a Peer To Peer (P2P) network, see figure 7.1.

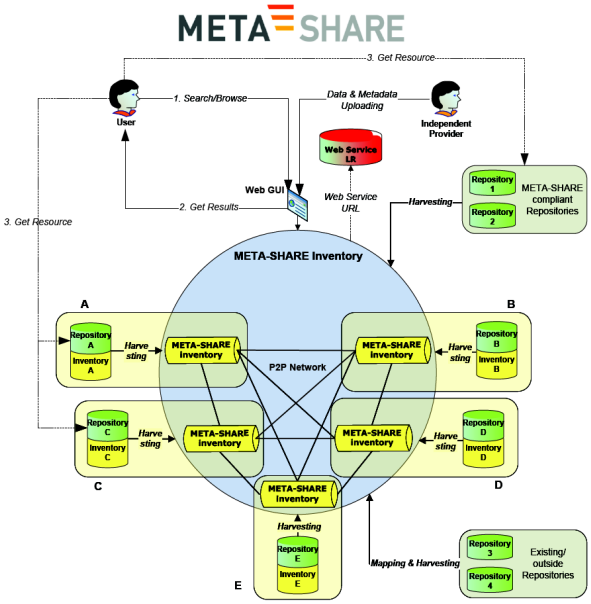

Figure 7.1: METASHARE system architecture.

Truly speaking, the architecture is more a P2P-like than a pure P2P. In fact each "core member", the members which are in the inner (yellow) circle in figure 7.2, offers the same services. This implies that, from the user point of view, accessing METASHARE means accessing a virtual "web portal" which provides a list of services, regardless of which member is offering the service.
Where the P2P aspects are in this architecture? Criteria are the following:

Each single node (member) is able to offer some services and/or manage some "piece of data";

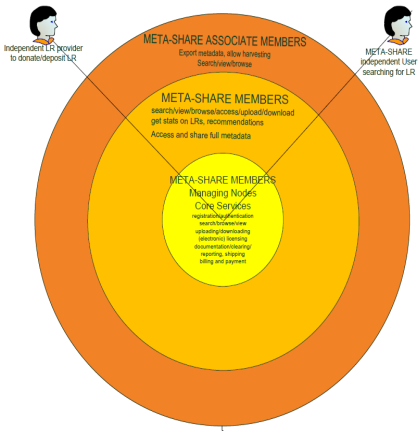

Figure 7.2: METASHARE growth rings.

Each service is replicated on each node;

Each "piece of data" is replicated on each node and then synchronized among nodes, so that each node will manage the complete set of data. One example being the catalog of offered Language Resources: even if each node has its own catalog, this one is synchronized among all nodes until a complete catalog is created. The complete catalog is, then, synchronized among the network so that each node will offer the complete catalog, see figure 7.3.

7.3 Towards a different architecture

The architecture depicted in figure 7.1 has been released for the first version of METASHARE[4]. Currently the NETwork for the Multilingual Europe Technology Alliance (METANET) consortium is working on a slightly different architectural model for the new release of METASHARE[5]. This new model plans to move from a P2P model to a sort of "Primus Inter Pares" one. In this model, each member manages some core services[6] and the virtual access point to the METASHARE

[4]METASHARE version 0, May 2010
[5]METASHARE version 1, May 2011
[6]These services represent the "core business" for members. For example the billing service, managed by ELRA, can be managed by ELRA itself in METASHARE.

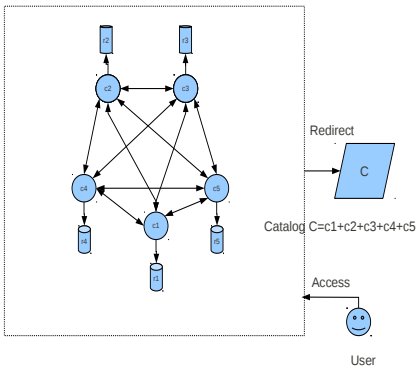

Figure 7.3: METASHARE virtual web portal and single catalog offered.

platform is able to redirect the request to the correct node offering the service, see figure 7.4.

7.3.1 Proposed architectures: Common and Distinctive Aspects

Architectures depicted in figures 7.1 (and in its simplified version 7.2) and 7.4 have some aspects in common as well as some distinctive characteristics, as briefly reported below.

Common Aspects In both architecture, METASHARE offers a virtual web portal to the community. This means that each member provides its own copy of the web portal. Each single IP address is mapped on a unique DNS entry, [45]. When the user joins the web portal[7] (s)he is redirected, by load balancing, to one of the physical portal provided by members. In addition, the architecture in version 1 allows that services can be shared among members (for instance, the service "Service 1" in figure 7.4);

Distinctive Aspects In the first version (version 0), each member has a copy of the whole set of services offered by METASHARE: this means that, regardless of which member the user is connected to, (s)he is able to access all

[7] http://www.meta-share.eu

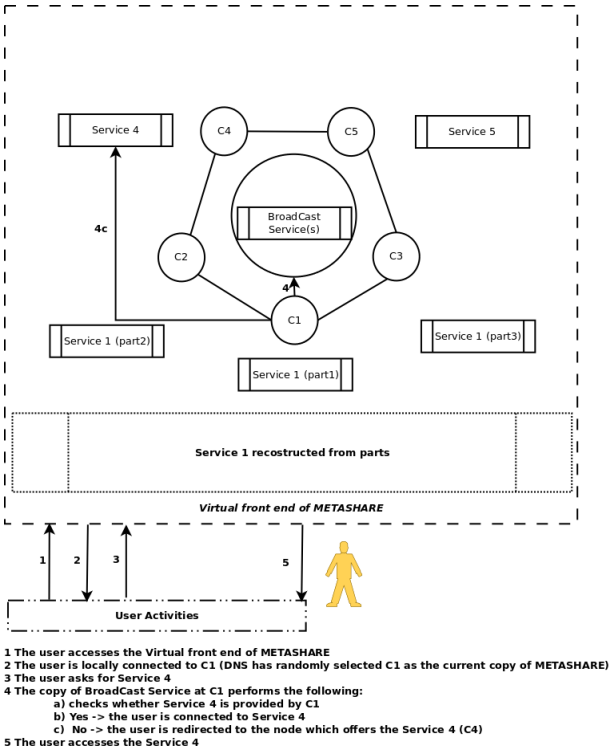

Figure 7.4: METASHARE virtual web portal and single service access.

provided services as they were *local*[8] ones. ON the contrary, in version 1, each node has a broadcast service which is in charge of:

a checking whether the user is actually requesting a *local* or a *remote*[9] service;

b checking whether the user has the rights to use the service, see section 7.4.1;

c redirecting the user to the member where the requested service is available, if necessary.

Figure 7.5 details the points **a** and **c** described above. In the simple scenario reported, the user joins the METASHARE web portal asking for "service 4" and

[8] *Local* services are services installed at the joined member.
[9] *Remote* means that the requested service is accessible from a different member.

(s)he is redirected by DNS load balancing to node (member) 1. The local copy of the "BroadCast Service" sends the user request to node (member) 4, where the "Service 4" is actually provided.

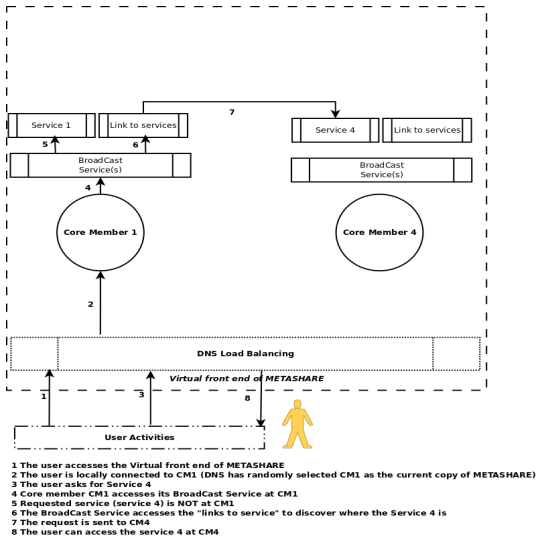

Figure 7.5: Broadcast service.

7.4 User management in METASHARE

The user management is still embryonic in METASHARE even if this topic is considered a fundamental one by the METANET consortium.
In fact, METASHARE, as a community, will be joined by many researcher, users, developers, simple curious (guests) as so on. All these kinds of users need to be managed in a way or another.
The user management is a mix of Identity Management to identify users, Access Management to create Access Control Lists on profiled users and, finally, to move the profiled identities at Language Resources level, to map the authorized users on the offered resources (Resource Management). The METANET consortium is working on such topic for METASHARE version 1 and it is currently analyzing (two) different solutions which are in relation with the (two) proposed architectures.

Distributed Solution This solution is related to the architecture of version 0 (fig. 7.1). Each node is able to register and manage users; these identities are then shared among all nodes. This solution has both advantages and disadvantages:

Advantages The main advantage is that the incoming users can join any node and be immediately recognized as a registered users. Users can join the METASHARE web portal via the DNS load balancing as well as by going directly to the portal offered by one node[10] and be sure to be recognized and granted access. In addition, any changes to the profiles of registered users can be performed in any node, see figure 7.6.

Disadvantages The synchronization of user's data is a critical point. The consortium has to guarantee that users can change any attribute of their profiles and this change is replicated to all nodes very fast. In fact, there are attributes that are critical for Identity Management (IM) and Access Management (AM) processes: the password needs to be synchronized to grant access from any node; attributes such as the membership[11] and affiliation[12] are strictly connected to ACLs and then to the actual access to language resources available in the infrastructure.

Local Solution This solution uses the broadcast service to check whether the incoming users have been created locally. In figure 7.7 the steps to authenticated are shown. User's data are not synchronized among nodes, this implies that the user can manage his(her) profile directly on the node where (s)he is registered. The attributes of the users are then always updated.

7.4.1 SSO in METASHARE

The Single Sign On in METASHARE is connected to the metadata issues in the METANET project. METANET has provided a first version of metadata set to describe resources and services. Since the consortium will apply the Combined Identity and Access Management strategy, (see section 6.2), the user's attribute, related to his(her) rights for accessing services and resources, are replicated at

[10]The virtual web portal being http://www.meta-share.eu, while the direct portal is, for instance, http://www.meta-share.it

[11]Membership is a concept derived from LDAP and identifies the groups to which users belong.

[12]Affiliation is related to the institute to which the users belong

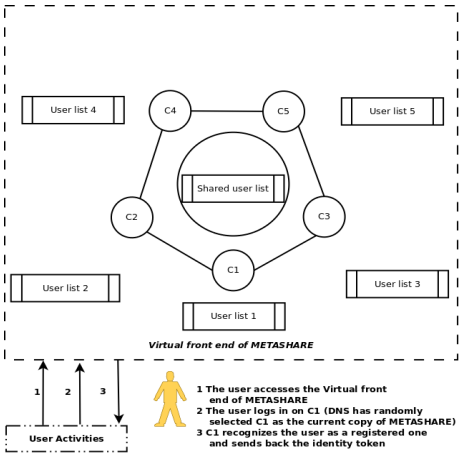

Figure 7.6: Users are shared among all members.

metadata level[13] as displayed in figure 6.1 and detailed in figure 7.8. After the user has been granted access to the METASHARE platform, his(her) rights for accessing resources and services are mapped onto the ones defined at resource level. If these set of attributes match, the user can access and use the resource (s)he asking to use. The process of accessing resources includes different processes defined in the linguistic infrastructure:

Resource Management to profile resources offered according to a list of access rights;

Single Sign On to allow the user join the platform with his(her) credentials;

CIMAM to check user's credential over resources *restrictive metadata*.

According to figure 7.8, the steps of the process are the following:

1. The user logs on the platform;
2. The user's credential are checked over the *restrictive metadata*;
3. If the user access to the requested resource is not granted, the user leaves the platform;
4. If the user access to the requested resource is granted, the user enters the platform to access the resource.

[13]This set of metadata is the *restrictive metadata* subset (see figure 2.1).

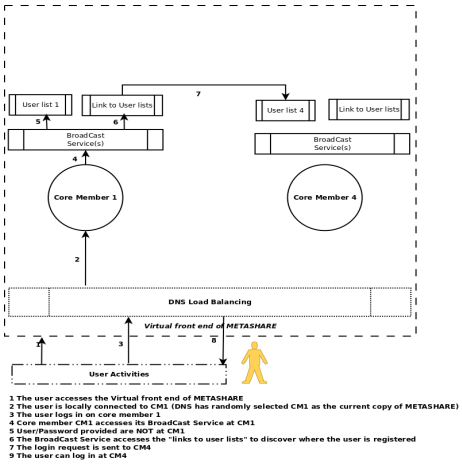

Figure 7.7: Users locally created at each node.

7.4.2 Implementations of Single Sign On in METASHARE

The interaction between the SSO and CIMAM processes described in previous section is independent from the actual implementation of SSO. Some attributes of the users connected to the rights for accessing resources and services, such as ACLs, memberships, affiliation ..., can be defined without an effective implementation of SSO in METASHARE. One possible solution is to use the SSO modules which come with the *Django* framework[14] such as *Oauth*[15].To use other techniques, such as *openid*[16] and Shibboleth is currently under discussion.
METASHARE is a "monotonically crescent distributed network of repositories" and the choice of Shibboleth is particularly suitable since this solution offers an easy way of sharing identities among the network and it will be very strategic when the network of METASHARE will start to increase.
In addition, future versions of METASHARE can gain benefits by implementing the Where Are You From (WAYF) service. This service is a native way to redirect users to their home institutions; following figure 7.7, the WAYF service can be used to redirect the user to the node where (s)he has been registered.

[14]The web portal http://www.meta-share.eu has been developed using Django -see http://www.djangoproject.com -
[15]http://oauth.net/
[16]http://openid.net/

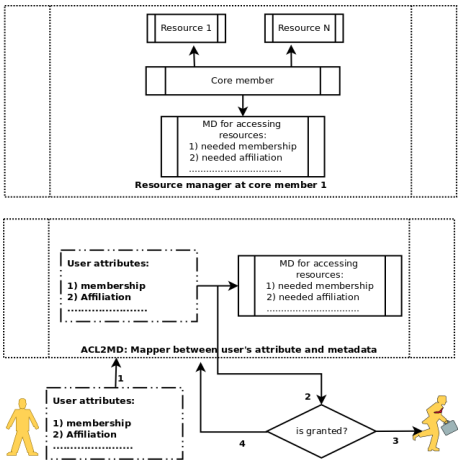

Figure 7.8: Metadata and user attributes mapping workflow.

7.4.3 SSO in PANACEA

The SSO in Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies (PANACEA) has not been analyzed yet. The project is more dedicated in implementing the basic infrastructure than in identity management. However, since PANACEA will provide secure web services to the industrial community, can not avoid a SSO system to protect the platform.

7.5 Security in LRIs

The security aspect is fundamental for Language Resource Infrastructures and it is strictly connected to Single Sign On. We have seen in section 7.1 that the SAML protocol is responsible for managing security assertions in a distributed infrastructure, so it should be used as security layer.
We have analyzed the security issues in both METANET and PANACEA and recognized that both projects present many security levels:

- Filter the access of the users. When a user joins the Language Resource Infrastructure (LRI), the user is enriched with a set of rights that will filter what a user can do. This is an aspect which is managed by IM and AM;
- Security at Service Provider level. This security is embedded with the SOAP envelope of the web services. Each web service available in the LRIs needs to be extended via a WSS4J[17] implementation;
- Shibboleth. We have tested the "shibbolization" (see appendix A of web services for the IDEntity Management federation.

We decided to work on both the WSS4J and Shibboleth technologies for managing the security in web services environment.
We re-deploy our web services including security directives according to the WSS4J guidelines. This solution is particularly suitable when the SAML protocol has been selected as security layer[18] of the infrastructure. Previously solution is specifically used for new web services, while the use of Shibboleth is suitable for managing security in consolidated environments.
We have used Shibboleth to protect web services already available at our site.

[17] http://ws.apache.org/wss4j/
[18] WSS4J uses SAML token as a security token.

Chapter 8

Afterword

As a child, I liked to play with Lego[1]. I was used to open the box, spread the colorful interlocking plastic bricks, all the array of gears, mini-figures and various other parts on the table and start building something that was not (obviously) on the cover of the box.
What was fascinating me, was the pretty much infinite number of combinations coming out from the simple fact of connecting one brick to another.
Well, many outcomes were nonsense, but someone of them looked really interesting. Even if I did not follow instructions, my works did seem to work. To have sort of "meanings".
Moreover, when I started building an object with specific features, it returned back to have different characteristics from the planned ones, but equally interesting: simply because rules and needs went changing during the development of the object.
Planning and building Hardware and Software infrastructures is like playing with Lego. Modules such as Authentication and Authorization, licensing management, registration, service provisioning and registration and "concepts" such as synchronization procedures, redundancy, high availability play the role of the plastic bricks: they can be arranged in many ways to create different infrastructures according to various requirements.

[1] Lego is trademarked as LEGO®

Chapter 9

Conclusion

In this thesis we have described our efforts in analyzing complex Language Resource Infrastructures.
In the past three years we have studied many typical building blocks of LRIs, including Identity Management, Access Management, security ... and we have studied how to personalize these building blocks according to different linguistic scenarios. We have been lucky since Europe projects where we are currently involved represent two well distinct LRIs, giving us the possibility of starting from general principles up to their *actual* implementations in real frameworks.
We have deeply presented the solution that we have implemented at Istituto di Linguistica Computazionale (ILC), showing that the modeling principles used in such implementation can be applied in more big platforms such the ones delivered in NETwork for the Multilingual Europe Technology Alliance (METANET) and Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies (PANACEA).
We recognized that there is many work to do, yet, to finalize the above projects about all possible issues that such infrastructures can cause. The most important is related to the Combined Identity and Access Management (CIMAM) that will be adopted in METASHARE. So far METASHARE is a pool of few partners which are interconnected and contain replicated information about Language Resource and Technology (LRT) available in the platform as well as their documentation and, finally, users registered in the infrastructure.
METASHARE aims to be a peer-to-peer-based infrastructure, but how it is cur-

rently designed, it is a distributed and replicated infrastructure. The idea for new version of METASHARE is to move towards a real peer-to-peer architecture, where information is de-localized and not simply replicated. We will use the Identity Management to start studying this architectural upgrade: we propose that each node of the platform will have its Authentication and Authorization Infrastructure (AAI) module but users are not synchronized among nodes. This means that the when the user joins the platform, (s)he creates his/her account on the node where is (locally) connected, and the next time the same user joins the framework the credentials (s)he provides are checked against the user created locally. If the user is present in the list of local identities, then (s)he is granted to the platform. On the contrary, the user's credential are broadcast to every node so that (s)he can log on the node where (s)he registered the account.
We believe that this solution will change the architecture from the current Peer To Peers (P2Ps)-like to a Primus Inter Pares one which is much closer to a real Peer To Peer model.

Chapter 10

Acronyms

AAI Authentication and Authorization Infrastructure Authentication and Authorization Infrastructure is a complex system which merges typical authentication issues such as simple login, single sign on, identity assertions ... with authorization topics such as rights on resources, copyright protected material ... 70

ACL Access Control List

ACL-HLT Association for Computational Linguistics: Human Language Technologies 8

AG Annotation Graph 35

AI Artificial Intelligence 5

AM Access Management 61

API Access Program Interface

CA Certification Authority 86

CAS Common Analysis Structure 42

CES Corpus Encoding Initiative 12

CIMAM Combined Identity and Access Management 69

CLARIN Common Language Resource Infrastructure Network 7^{th} Framework Program; Capacities Specific Program; Research Infrastructures: Grant agreement no.: 212230 21

CMDI CLARIN Meta Data Infrastructure 36

COLING Conference on Computational Linguistics 7

DC Dublin Core ISO Standard 15836, and NISO Standard Z39.85-2007. 36

DCMI Dublin Core Meta Data Initiative 12

DCR Data Category Registry 80

DFKI Deutsche Forschungszentrum für Künstliche Intelligenz The German Research Center for Artificial Intelligence 8

EAMT European Association for Machine Translation 8

ELDA Evaluation and Language Distribution Agency ELDA is a company responsible for managing/selling Language Resources 6

IdP Identity Provider 85

IJCNLP International Joint Conference on Natural Language Processing 8

ILC Istituto di Linguistica Computazionale 69

IM Identity Management 61

IMDI ISLE Meta Data Initiative 36

Interspeech International Speech Communication Association 8

IPR Intellectual Property Rights 37

IR Information Retrieval 25

LDC Linguistic Data Consortium According to the LDC website[1] "the Linguistic Data Consortium supports language-related education, research and technology development by creating and sharing linguistic resources: data, tools and standards." 6

LE Language Engineering 34

LG Language Grid 29

LRE Language Resources and Evaluation 7

LREC Language Resources and Evaluation Conference 7

[1] http://www.ldc.upenn.edu

LIRICS Linguistic Infrastructure for Interoperable Resources and Systems Project No.22236 - 2005-2007 18

LMF Lexical Markup Framework LMF, ISO-24613:2008, has been revised up to revision 16 19

LR Language Resource Generic Term for identifying a language resource related to *data*. See section 2.1 for detailed information. 44

LRI Language Resource Infrastructure Generic term for identifying a language resource infrastructure. A language resource infrastructure is a complex organism composed by different components, see chapter 5. 86

LRT Language Resource and Technology Generic term for identifying both a dynamic and a static language resource. See section 2.1 and its subsections for detailed information. 69

LT Language Technology Generic Term for identifying a language resource, usually a tool, used to manage data Language Resources. See section 2.1 and its subsections for detailed information. 25

LTp Language Technologies

MAF Morphological Annotation Framework MAF, ISO/DIS 24611 (Under development) 19

METANET NETwork for the Multilingual Europe Technology Alliance This project is also known as Technologies for the Multilingual European Information Society (T4ME) and is funded by the European Commission through the Seventh Framework Program, Grant agreement no.: 249119 69

MPEG Moving Picture Experts Group 9

MT Machine Translation 23

NICT National Institute of Information and Communications Technology 29

NLP Natural Language Processing NLP is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages [46] 38

NLSR Natural Language Software Registry 9

OAI Open Archive Initiative

OAI-PMH Open Archive Initiative Protocol for Metadata Harvesting

OLAC Open Language Archives Community 12

ORI Open Resource Infrastructure 26

PANACEA Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies 7^{th} Framework Program; Information and Communication Technologies Grant agreement for: Small or medium scale focused research project(STREP). Grant agreement: 248064 69

PID Persistent IDentifier Persistent identifier is a unique identifier that some systems define for generic resources. ISBN is a persistent identifier. 23

P2P Peer To Peer 70

PIP Primus Inter Pares

RM Resource Management 50

RnD Research and Development

SAML Security Access Markup Language 85

SME Small and Medium Enterprise 17

SP Service Provider 85

SSO Single Sign On 55

SynAF Syntactic Annotation Framework ISO 24615:2010 (Published) describes the syntactic annotation framework (SynAF), a high level model for representing the syntactic annotation of linguistic data. 19

T4ME Technologies for the Multilingual European Information Society 75

TEI Electronic Text Encoding Interchange 9

TO Traveling Object The traveling object is an exchange format developed in the PANACEA project. Details can be found in [24]. 25

UC Universal Catalog 6

UIMA Unstructured Information Management Architecture 39

VLO Virtual Language Observatory 8

WAYF Where Are You From 86

Bibliography

[1] N. Calzolari, "Community culture in language resources - an international perspective. 2. towards a research infrastructure for language resources," in *Workshop in conjunction with LREC*, 2006, pp. 12–15.

[2] A. Zampolli, "Towards reusable linguistic resources," in *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*, ser. EACL '91. Stroudsburg, PA, USA: Association for Computational Linguistics, 1991, pp. 1–1. [Online]. Available: http://dx.doi.org/10.3115/977180.977181

[3] K. Choukri, V. Arranz, V. Mapelli, H. Mazo, D. Mostefa, and N. Moreau, "Up-to-date chart of lr and players and classification along different lines (deliverable d2.1a)," Flarenet, Tech. Rep., 2009.

[4] N. Calzolari, C. Soria, N. B. Valeria Quochi, G. Budin, T. Caselli, K. Choukri, J. Mariani, M. Monachini, J. Odijk, and S. Piperidis, "Flarenet blueprint of actions and infrastructures (deliverable d8.2b)," Flarenet, Tech. Rep., 2010.

[5] N. Calzolari, C. Soria, N. Bel, G. Budin, K. Choukri, J. Mariani, M. Monachini, J. Odijk, S. Piperidis, V. Quochi, and A. Toral, "Flarenet blueprint of actions and infrastructures (deliverable d8.2a)," Flarenet, Tech. Rep., 2009.

[6] M. Gavrilidou, P. Labropoulou, S. Piperidis, M. Speranza, M. Monachini, V. Arranz, and G. Francopoulo, "Specification of metadata-based descriptions for language resources and technologies (deliverable d7.2)," METANET, Tech. Rep., March 2011, dissemination Level: Public.

[7] N. Calzolari, C. Soria, R. D. Gratta, S. Goggi, V. Quochi, I. Russo, K. Choukri, J. Mariani, and S. Piperidis, "The lrec map of language resources

and technologies," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010.

[8] T. Consortium. (2010, January) Tei p5: Guidelines for electronic text encoding and interchange. [Online]. Available: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

[9] N. Calzolari, P. Baroni, and C. Soria, Eds., *Language Resources of the future the future of Language Resources.* Barcelona: The European Language Resources and Technologies Forum, 11-12 February 2010.

[10] D. Broeder. (2010, August) Clarin metadata and iso Data Category Registry (DCR). [Online]. Available: www.isocat.org/2010-TKE/presentations/CMDI-ISOCat.ppt

[11] Keith G. Jeffery . (2009, September) Metadata in the european e-infrastructure. [Online]. Available: http://www.csc.fi/english/pages/neeri09/programme/materials-thu/jeffery2.pdf

[12] C. Federmann, B. Georgantopoulos, Riccardo del Gratta , B. Magnini, D. Mavroeidis, S. Piperidis, and M. Speranza, "Meta-share functional and technical specification (deliverable d7.1)," METANET, Tech. Rep., January 2011, dissemination Level: Restricted.

[13] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence)*, 1st ed. Prentice Hall, 2000, neue Auflage kommt im Frhjahr 2008.

[14] N. Calzolari, P. Baroni, N. Bel, G. Budin, K. Choukri, S. Goggi, J. Mariani, M. Monachini, J. Odijk, S. Piperidis, V. Quochi, C. Soria, and A. Toral, Eds., *Shaping the Future of the Multilingual Digital Europe.* Vienna: The European Language Resources and Technologies Forum, 12-13 February 2009.

[15] G. Francopulo, J. Nioche, and M. Kemps, "Evaluation of existing standards for nlp lexica (deliverable d2.1)," Lirics-loria, Tech. Rep., 2005.

[16] M. Monachini, C. Soria, M. Ulivieri, N. Calzolari, T. Declerck, and M. Mammini, "Guidelines and tools for producing standards, test-suites and api(s) (deliverable d1.1)," Lirics-loria, Tech. Rep., 2005.

[17] A. Funk, K. Bontcheva, N. Aswani, I. Roberts, and J. Nioche, "Api for morpho-syntactic annotations (deliverable d5.1 c v2)," Lirics-loria, Tech. Rep., 2007.

[18] J. Nioche, K. Bontcheva, A. Funk, N. Aswani, and I. Roberts, "Api for syntactic annotations (deliverable d5.1 d v2)," Lirics-loria, Tech. Rep., 2007.

[19] D. Heimbigner and D. McLeod, "A federated architecture for information management," *ACM Trans. Inf. Syst.*, vol. 3, no. 3, pp. 253–278, 1985.

[20] D. van Uytvanck. (2009, February) Persistent identifier service. [Online]. Available: http://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf

[21] M. Kemps-Snijders, S. E. Wright, and M. Windhouwer. (2010, February) Isocat a short introduction. [Online]. Available: http://www.clarin.nl/system/files/ISOcat-20100208.pdf

[22] D. van Uytvanck. (2009, February) Persistent identifier service. [Online]. Available: http://www.clarin.eu/files/centres-CLARIN-ShortGuide.pdf

[23] K. Lindén. (2010, February) Initial clarin service provider federation. [Online]. Available: http://www.clarin.eu/files/SPF-CLARIN-ShortGuide.pdf

[24] M. Poch, P. Prokopidis, G. Thurmair, C. Schnober, R. Del Gratta , N. Bel, and O. Hamon, "Architecture and design of the platform (deliverable d3.1)," PANACEA, Tech. Rep., 2010.

[25] T. Ishida, "Language grid: An infrastructure for intercultural collaboration," in *In IEEE/IPSJ Symposium on Applications and the Internet SAINT-06*, 2006, pp. 96–100.

[26] Language Grid Team . (2010, January) Language grid pamphlet. [Online]. Available: http://langrid.nict.go.jp/file/LanguageGrid20101117en.pdf

[27] N. Calzolari, "Approaches towards a lexical web: the role of interoperability," in *ICGL 2008: The First International Conference on Global Interoperability for Language Resources*, Hong Kong, SAR, January 2008, pp. 34–42.

[28] Language Grid Team . (2006, May) Language grid: Connecting world's language services to support intercultural collaboration. [Online]. Available: http://langrid.nict.go.jp/file/langrid20060501e.PDF

[29] ——. (2010, August) Language grid services. [Online]. Available: http://langrid.nict.go.jp/file/langridservice20100820en.pdf

[30] U. Heid, H. Schmid, K. Eckart, and E. Hinrichs, "A corpus representation format for linguistic web services: The d-spin text corpus format and its relationship with iso standards," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010.

[31] D. Ferrucci and A. Lally, "Uima: an architectural approach to unstructured information processing in the corporate research environment," *Nat. Lang. Eng.*, vol. 10, no. 3-4, pp. 327–348, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?id=1030318.1030325

[32] H. Cunningham, "Gate, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, pp. 223–254, 2002.

[33] S. Bird and M. Liberman, "A formal framework for linguistic annotation," *Speech Communication*, vol. 33, no. 1-2, pp. 23–60, 2001.

[34] N. Ide and K. Suderman, "Bridging the gaps: Interoperability for graf, gate, and uima," in *Proceedings of the Third Linguistic Annotation Workshop*. Suntec, Singapore: Association for Computational Linguistics, August 2009, pp. 27–34. [Online]. Available: http://www.aclweb.org/anthology/W/W09/W09-3004

[35] ——, "Graf: A graph-based format for linguistic annotations," in *Linguistic Annotation Workshop, ACL 2007*, Prague, 2007.

[36] D. van Uytvanck. (2009, February) Component metadata. [Online]. Available: http://www.clarin.eu/files/metadata-CLARIN-ShortGuide.pdf

[37] N. Calzolari, "European initiatives to promote cooperation between speech and text communities," in *INTERSPEECH*. ISCA, 2004.

[38] *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*. ISCA, 2004.

[39] A. P. Sheth and J. A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Comput. Surv.*, vol. 22, no. 3, pp. 183–236, 1990.

[40] Riccardo Del Gratta , R. Bartolini, T. Caselli, M. Monachini, C. Soria, and N. Calzolari, "Ufra: a uima-based approach to federated language resource architecture," in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias, Eds. Marrakech, Morocco: European Language Resources Association (ELRA), may 2008, http://www.lrec-conf.org/proceedings/lrec2008/.

[41] O. Foundation. (2007) Open ldap. http://www.openldap.com/.

[42] M. Erdos and S. Cantor. (2001) The Shibboleth architecture. http://shibboleth.internet2.edu/.

[43] Sue Ellen Wright, "A global data category registry for interoperable language resources." in *Proceedings of LREC 2004*, ELRA, Lisbon., 2004.

[44] J. Ganci, S. Cherukuwada, B. McGoogan, L. Olivera, J. Thorwart, and W. Wardell. (2010, August) Identity and access management solutions using websphere portal v5.1, tivoli identity manager v4.5.1 and tivoli access manager v5.1 (redbook). [Online]. Available: http://langrid.nict.go.jp/file/langridservice20100820en.pdf

[45] T. T. Guide. (2005, September) Dns name server load balancing. [Online]. Available: http://www.tcpipguide.com/free/t_DNSNameServerLoadBalancing.htm

[46] E. Charniak, *Introduction to artificial intelligence*. Addison-Wesley Professional, 1984.

Appendix A

Shibboleth implementation of Single Sign On

Shibboleth[1], is an open source middleware software that allows sites to make informed, federation-wide authorization decisions for individual access of protected online resources in a privacy-preserving manner. It is based on Security Access Markup Language (SAML), and largely adopted by research and academic communities. Shibboleth is a powerful framework (SAML2-based) able to guarantee a trusted communication and a single-sign-on within a federated architecture. Shibboleth, out of the box, is an Identity Provider (IdP) and a Service Provider (SP). The Identity Provider is based on a directory (LDAP is the most frequent choice, but database can serve as directories as well) and a web application which can be deployed under servlet containers.
In details, the directory (LDAP, for instance) must complain to at least the *eduPerson*[2] schema, with some additional attributes from other schemes which depend on the specific guidelines of the federation. For example the *schac*[3] scheme is needed for academic purposes. The IdP is configured to read attributes from the directory and send them to the SP via a secured HTTP session where SAML messaging are used to exchange attributes and values.
The Service Provider is, essentially, a daemon and it needs to be installed on each

[1] http://shibboleth.internet2.edu
[2] http://middleware.internet2.edu/eduperson/
[3] http://www.terena.org/activities/tf-emc2/meetings/7/slides/schac-20061016-tf_emc2-malaga.pdf

data center that offers services. IdPs and SPs communicate via backhand certificates (which can be self signed or from a Certification Authority (CA)) and share a metadata file. This file contains all descriptive metadata for each component that belong to the federation. The trusted communication is performed via the "relying party" shared file which works with the metadata file to check whether the trusted communication between IdPs and SPs can be performed. These two files allow people, which join the federation, to exchange their identities among all participants.

Language Resource Infrastructures designed to be network of repositories will gain from the choice of Shibboleth. It is particularly suitable, since offers an easy way of sharing identities among the network, while the two files (metadata and relaying party) described above serve as the contract between identities (users and providers), institutions (repositories) and offered services in the infrastructure. Figure A.1 shows the basic architecture of Shibboleth.

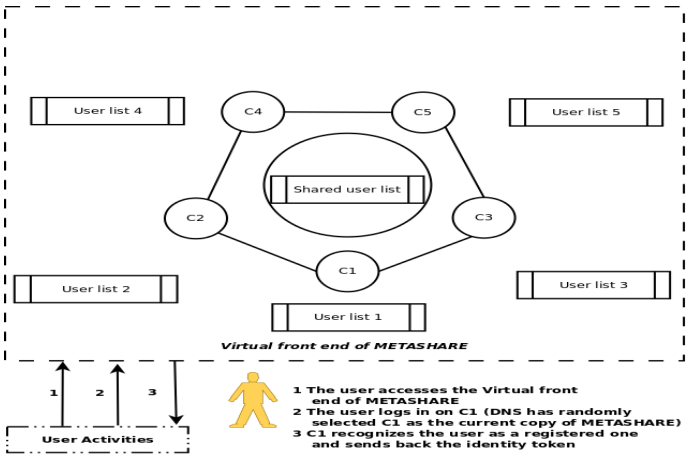

Figure A.1: Shibboleth concept.

In addition, Shibboleth offers the where are you from Where Are You From (WAYF)[4] service. This service plays the role of a "central" point of access, where users are redirected when access services.

Shibboleth can be used to protect services provided by the Language Resource Infrastructures (LRIs). This possibility is known as "shibbolize" an application. In other words, Shibboleth takes care of all identities and security issues. The

[4]http://www.wayf.dk/wayfweb/frontpage.html

simple version of such a protection is to cover an apache directory with Shibboleth. Each times the web application (whose end point is the protected apache directory) is accessed, the request is sent to Shibboleth that asks for the identities credentials, see figure A.2.

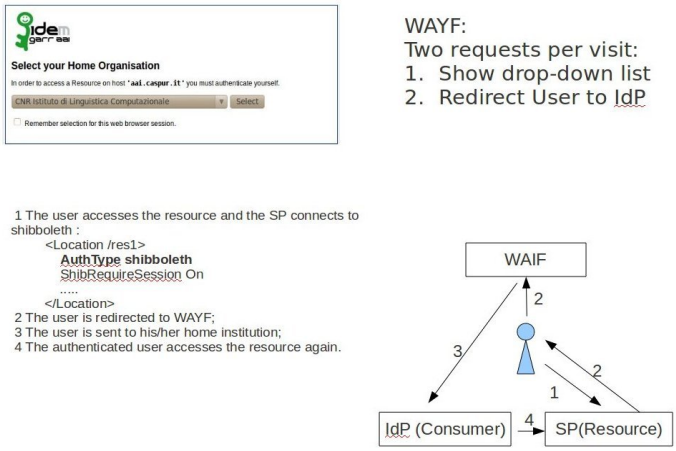

Figure A.2: Shibboleth concept with WAYF service implemented.

Figure A.2 shows the process of a user authentication done via a WAYF. Implementing the WAYF will be useful for managing largely distributed infrastructures, since is a native way to redirect users to their home institutions.