CLARIN and FLaReNet:

new European Initiatives

for Language Resources and Language Technologies

Nicoletta Calzolari

Istituto di Linguistica Computazionale del CNR, Pisa, Italy
glottolo@ilc.cnr.it

Today, many vitality & success signs… for LRs

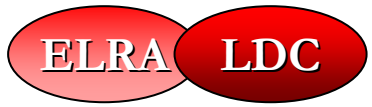
- In Spoken, Written, Multimodal areas … … in new emerging areas
- Statistical approaches…
- Different dimensions & layers: Content (Ontologies), Emotion, Time, …
- For Evaluation
- For Training
- …

- LREC (> 900 submissions); many LRs at **COLING** and even at **ACL**!!
- **ELRA** (self-sustaining) & **LDC**
- **LRE** (new Journal: N. Ide & NC)
- **ISO-TC37-SC4/WG4** (International Standards for LRs)
- **AFNLP**…
- **ESFRI - CLARIN** (also political & strategic role)
- **New calls** or **initiatives** in EU, US, ASIA, on LRs, interoperability, cooperation, …

BUT … an important point:

In the '90s

- There was a global vision of the field & its main components:
  - Standards
  - Creation of LRs
  - Distribution



Then:

  - Automatic acquisition

… towards the
Infrastructure of LRs & LT

While today:

- There is an ever increasing set of initiatives for new LRs, basic robust technologies, models??, algorithms,

- **We have a LR community culture**
- **BUT sort of scattered, opportunistic, not much coherence**

Today …

The wealth of data & of basic technologies is such that:

We should reflect again at the field as a whole & ask if

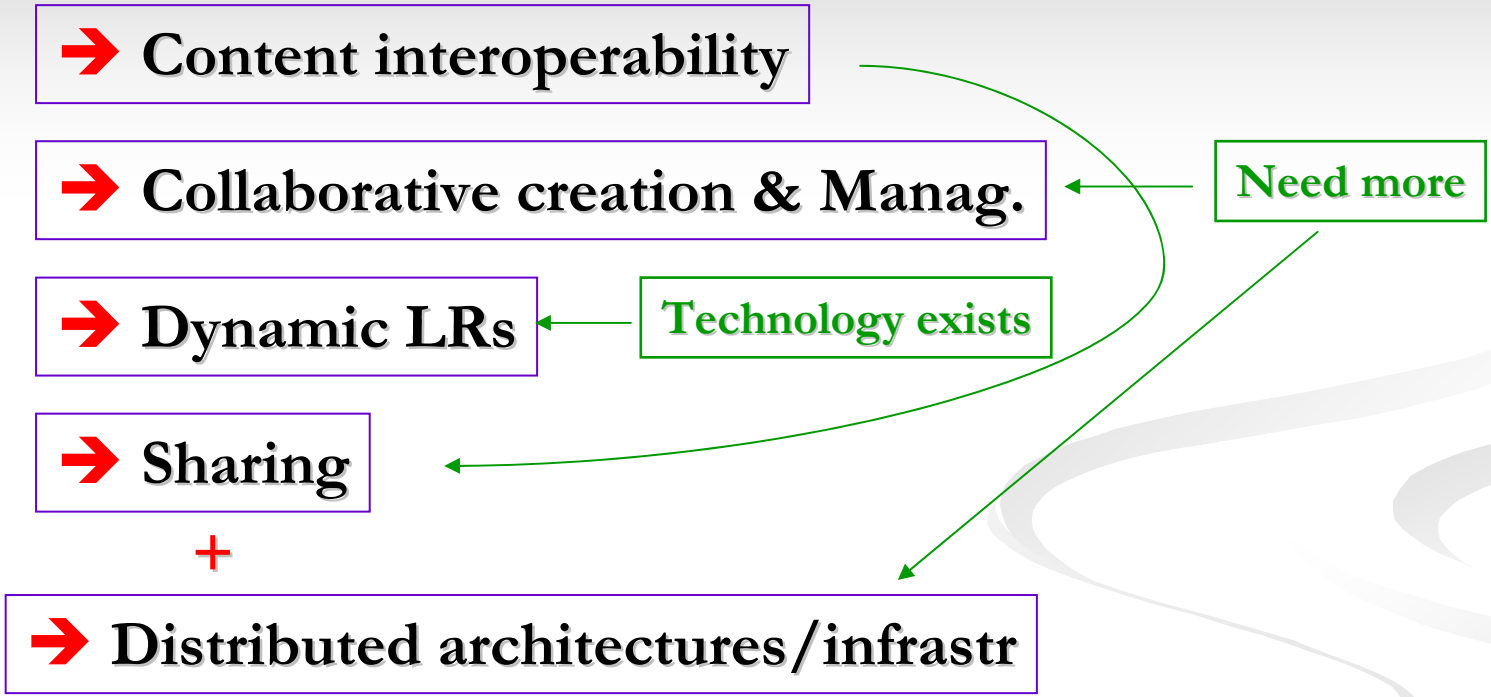
- **Standards** ← **Content interoperability**
- **Creation of LRs** ← **Collaborative creation & Manag.**
- **Automatic acquisition** ← **Dynamic LRs**
- **Distribution** ← **Sharing**

are still **"the"** important components,

or how they have changed/must change

… Which new challenges towards a new & more mature infrastructure of LRs & LTs??

These dimensions



could be at the basis of a new Paradigm for LRs & LT & of a new Infrastructure ??

In the "Semantic Web" vision ...

…need to tackle the twofold challenge of

- ***content availability &***
- ***multilinguality***

Natural convergence with **HLT**:

- ***multilingual semantic processing***
- ***ontologies***
- ***semantic-syntactic computational lexicons***

Builds also on EAGLES/ISLE

ISO & LIRICS:
LMF Meta-model + Data Categories

Objectives

- Design of the **abstract lexical meta-model**
- Definition of the **common set** of related **Data Categories**

LMF = an ISO standard for NLP lexica

- A **Lexical Markup Framework** as general & abstract meta-model & a set of structural nodes relevant for linguistic description
- A flexible environment enabling specific implementations of user-defined mark-up languages (called LML) on the basis of **common Data Categories**

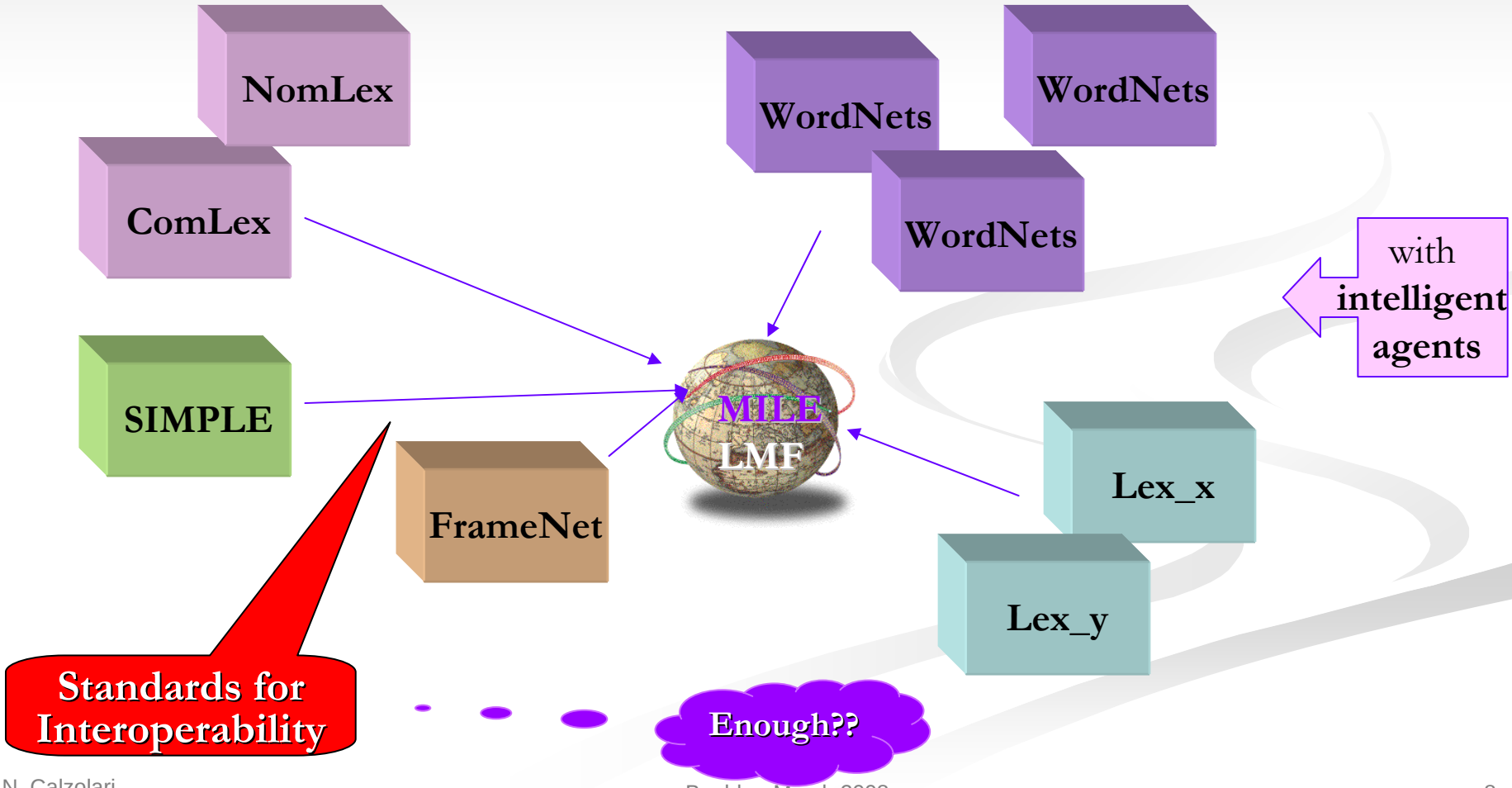
NEDO Asian Lang.

The field is mature

from Monica Monachini

Lexical WEB & Content Interoperability → 'Standards'

- As a critical step for **semantic mark-up** in the SemWeb



Need of tools to make this vision operational & concrete

New **prototype** **"LeXFlow":**

(http://xmlgroup.iit.cnr.it:98/MILE/lexflow/demo.xhtml)

- **web-based collaborative environment for semi-automatic management/integration of lexical resources**
- enabling interoperability of ***distributed*** lexical resources
- accessed by different types of ***agents***

From Language Resources

To ***Language Services***

Some steps for a "new generation" of LRs

- **From** huge efforts in building static, large-scale, general-purpose LRs
- **To** non-static LRs rapidly built on-demand, tailored to spefic user needs
- **From** closed, locally developed and centralized resources
- **To** LRs residing over distributed places, accessible on the web, choreographed by agents acting over them
- **From** Language Resources
- **To** ***Language Services***

Distributed Language Services

- A long-term scenario implying
  - content interoperability standards,
  - supra-national cooperation and
  - development of architectures enabling accessibility

- Create new resources on the basis of existing
- Exchange and integrate information across repositories
- Compose new services on demand

- **Collaborative & collective/social development and validation**, cross-resource integration and exchange of information

A new paradigm of R&D in LRs & LT

- **Open & distributed linguistic infrastructures for LRs & LT**
  - adopting the paradigm of **accumulation of knowledge** so successful in more mature disciplines, based on sharing LRs & tools
  - ability to build on each other achievements, results accessible to various systems, allowing controlled & **effective cooperation of many groups on common tasks** **(see HGP → HLP)**
- **Emerging concept of collective intelligence**
- Emphasize **interoperability** among LRs, LT & knowledge bases
  - e. g. initiatives aimed at achieving international consensus on annotation guidelines: to merge annotation efforts, produce coherent, comprehensive linguistic annotations to be readily disseminated throughout the community
- New ways of extending large-scale LRs and knowledge bases **relying on volunteer labour**, **wiki-mode?**

interoperability

Why an infrastructure of LRs?

Many dimensions around the notion of language

Political issues

e.g. a commonly agreed list of minimal requirements for "national" LRs: **BLARK**

Putting together technical, organisational, strategic, economic, political issues of LRs

Multilingualism

Cultural issues

➔Langua… …ral identity

…manities

Need of bodies for a broad research agenda & strategic actions for LT&LRs (W/S /MM)

Economic, soci… issues

➔Applications

➔Services

Interdisciplinarity & Multidisciplinarity

Technical issues

Technologies exist, but the infrastructure that puts them together and sustains them is still missing

Which Communities?

for

- Humanities
- Social Sciences
- Digital Libraries
- Cultural Heritage
- …

core

- Language Resources
- Language Technologies
- Standardisation

Enabling infrastr

on

- Grid
- Semantic Web
- Ontologists
- ICT
- …

Multilinguality

CLARIN ResInf

Focus on cooperation

for

- Many **application** domains (eculture, egovernment, ehealth, …)

CLARIN

Common Language Resources and Technologies Infrastructure

ESFRI Research Infrastructures

- Large-scale pan-European collaborative effort to make LRs and LTs available & readily useable to scholars of humanities & social sciences (& all disciplines)
- Need to overcome the present fragmented situation by harmonising structural and terminological differences
- Basis is a Grid-type of infrastructure and Semantic Web technology
- The benefits of computer enhanced language processing become available only when **a critical mass of coordinated effort is invested in building an enabling infrastructure**, which can **provide services in the form of provision of tools and resources as well as training and counseling** across a wide span of domains
- This is the mission of the CLARIN infrastructure initiative with a **EU wide collaboration (31+ countries)**
- The infrastructure will be based on a number of resource, service and expertise centres

How can we tackle these challenges?

J. Taylor
"eScience is about global collaboration in key areas of science and the next generation of infrastructures that will enable it"

Need to build new types of platforms

- to allow researchers to combine existing resources easily to new ones to tackle the big challenges
- to increase the productivity of all interested researchers, since currently too much time is wasted by preparatory work

from P. Wittenburg

eScience Vision

CLARIN is establishing such a new generation of extended infrastructure

Thus CLARIN is **not** about creating and building new language resources and technology, but

- making them available and accessible
- as services
- in a stable and persistent infrastructure

to allow tackling the great challenges

CLARIN:	http://www.clarin.eu
Grid Project:	http://www.mpi.nl/dam-lr
ISO TC37/SC4:	http://www.tc37sc4.org
Standards Project:	http://lirics.loria.fr/

from P. Wittenburg

Mission

- CLARIN will **create a comprehensive and free to use distributed archive of LRs & LTs** covering not only the languages of all member states, but also other languages studied and used in Europe

- Through the fact that the **tools and resources will be interoperable across languages and domains,** will contribute to addressing the issue of preserving and **supporting multilingual and multicultural European heritage**

- **An operational open infrastructure of web services will introduce a new paradigm of distributed collaborative development**

- It will allow **many contributors to add all kinds of new services** based on existing ones thus ensuring reusability and allowing scaling up to suit individual needs

We have still a long path …

& also a "new project":

in an **e-Content*plus*** Call for a:

- **"Thematic Network on Language Resources":**

FLaReNet

- **To provide common recommendations (to the EC) for future actions**
- **To give priorities**
- **Need of 'visions'**

In a global context, in cooperation with *CLARIN*
& also with *non-EU members*