Open Language Resources & Meta-Resources: a Treasure and a Challenge for Linked Data

The challenges of openness, interoperability, collaboration, …

Nicoletta Calzolari

ILC – CNR & ELRA

glottolo@ilc.cnr.it

www.flarenet.eu

NLP is "data intensive"

Story about

I prefer
Language Resources

Infrastructural nature

- Like railway, electricity, …

i.e. the backstage
Not in the forefront wrt applications

Infrastructural issues

Here in the forefront

Also some Calls for Cooperation

FLaReNet Recommendations: a global perspective

International Cooperation

INFRASTRUCTURE

Sustainability

Recognition

Develop ment

Docume ntation

Interoper ability

Availa bility

Cover age

Resource Interoperability

"Design and set up an interoperability framework for LRT"

- **Facts**
  - Essential **prerequisite for successful data exploitation**
  - **Lack** of interoperability and compliance with standards **costs a fortune**
    - While there is still the attitude ***"Why should I care?"***
- **Actions to be taken**
  - **Invest more** in standardisation activities
  - Make **standards operational** and put them in use
  - **Encourage/enforce use of best practices/**standards in LR production
  - Identify new **mature areas** for standardisation & **promote joint efforts**, also **among communities**

Danger of re-doing ...

→ **RDF conversion for LRs in LLOD**

Also in the LREC Challenge

META-DATA

Resource Documentation

"Ensure that LRs are accurately and reliably documented"

- **Facts**
  - LRs are often poorly documented or not documented at all
  - The gateway to discovery of LRs: **non documented LRs don't exist**
  - Helps understand, replicate, evaluate data: it **allows resource reusability**

RE Map & META-SHARE

- **Actions to be taken:**
  - Ensure that **appropriate metadata** are consistently adopted
  - Set up a global infrastructure of common **interoperable** metadata sets
  - Devise and adopt a widely → **agreed standard documentation template** for each resource type, based on identified best practice(s)

Call for action/ cooperation

Rationale

Accumulation of **massive amounts** of
- **multi-dimensional data** &
- **meta-data**

is a key to foster advancement

The **history of LRs** brings us through concepts such as
- **Reusability**
- **Integration**
- **Standards and Interoperability**
- **Cooperative projects**
- **Subsidiarity**
- **Infrastructural role of LRs**
- **Sharing**
- **…**

LRs → **Natural evolution** → **LRs & Metadata building as a collaborative "shared task"**

META-DATA

Conferences & LRs in the LRE Map

- LREC2010
- COLING2010
- ACLHLT2011
- IJCNLP2011
- Interspeech2011
- LTC2011
- RANLP2011
- O-COCOSDA2011
- LREC2012
- COLING2012
- NAACL2013
- Interspeech2013
- RANLP2013

- LREC2014
- COLING2014
- ACL2014
- ...

http://www.resourcebook.eu/

About 7,000 LR entries
Simple set of Metadata

- **Normalisation of LR names** and other values, through auto-completion

Community built

Temporal, usage, .. dimensions

Soon exported in LLOD

META-SHARE architecture

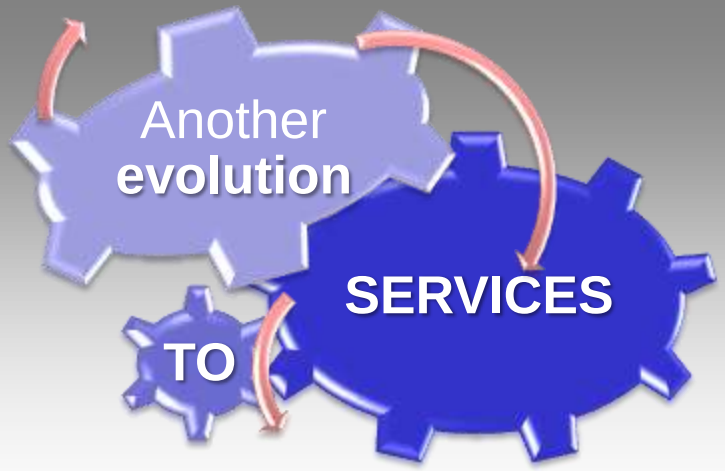
- **LRs** and their **metadata** (MD) reside in the **local repositories**
  - Rights of use and related restrictions under the control and responsibility of LR owners and the repository where the LR resides
- Each repository
  - maintains an inventory (a local inventory) with all MD of their LRs
  - exports MD
  - allows their harvesting.
- Harvested MD are stored in the META-SHARE central servers, with **synchronised inventories at all times**
- Central servers create, host and maintain a central inventory with all MD descriptions of all LRs available in the distributed network.

Reference model for LR development
Quantity, Quality & Adequacy to technological purposes

- **Facts**
  - In current **data-driven paradigm, innovation** depends on **big amounts of data**, of the **right type,** appropriate **quality,** & for **many languages**
- **Warning**
  - Dependence on data may create **disparity** for **under-resourced languages** and domains
- **Actions to be taken**
  - **Go Green**: enforce recycling, reusing and repurposing
  - **Invest in Web 2.0/3.0 methods for LR collaborative creation**

E.g. with LLOD

USE



LRs as services &
Services around LRs

- **LRs as services**
  - Composite access
  - Web-services for Visualisation, Analysis, ...
  - Extracting, Adapting, Merging, Linking, ...
  - …

USE

Services

- Services around LRs
  - Inventorying
  - Describing: with Metadata
  - Sharing: Authentication, ...
  - Legal: licensing, …
  - Web-services for Collecting, Crawling, Cleaning, Accessing, Linking, Integrating, Clustering, ...
  - Converting (around Interoperability): e.g. in RDF
  - Annotating , (Content) Analysing, Acquiring info, ...
  - Adapting , Repurposing, Evaluating,
  - Crowdsourcing
  - Translating, Localising, ...
  - Summarising , Mining, ...
  - Understanding, ...
  - …

From language neutral ...

Towards more language specific (*basic and advanced*)

And language-based

Call for action/ cooperation

USE

Distributed Language Services

A scenario implying:

- **content interoperability standards**
- **international cooperation**
- **architectures enabling accessibility**

Enabling:

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

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

SHARE your LRs

USE

o enable reuse, replicability of experiments, …

To start sharing the linguistic knowledge the field is able to produce

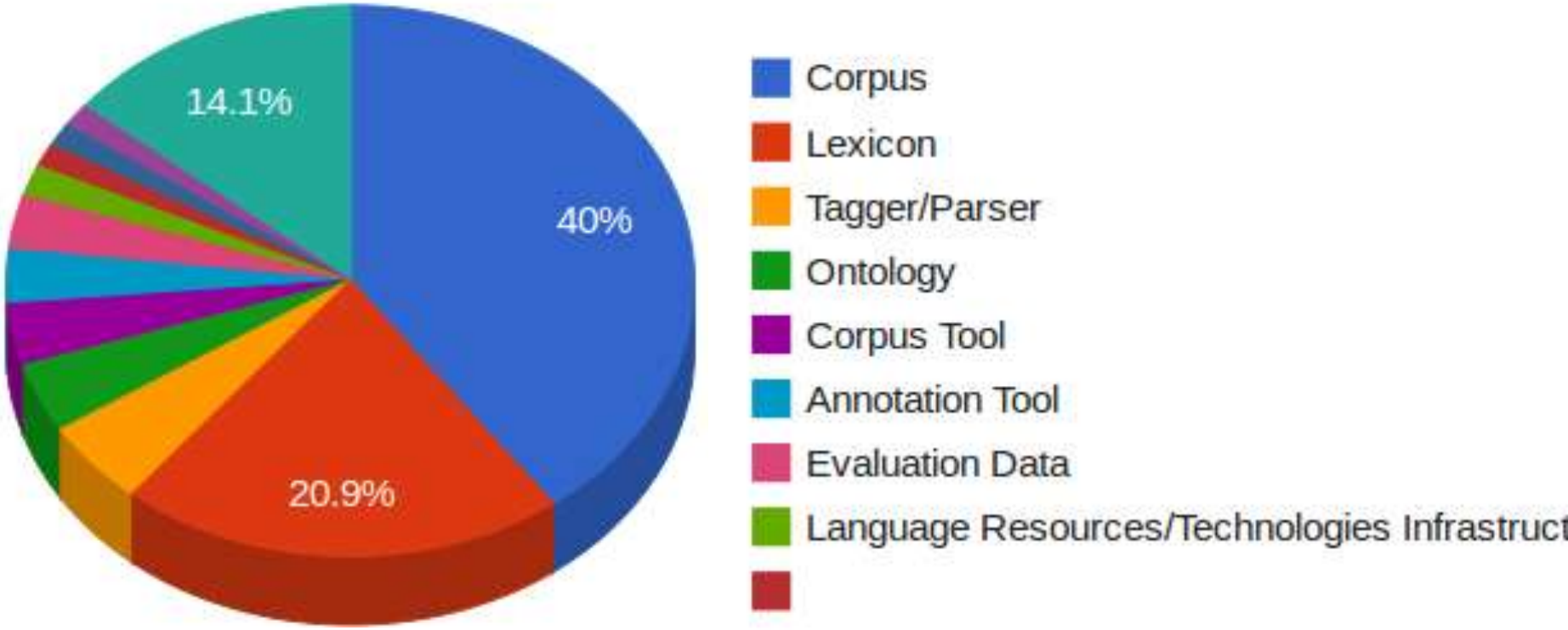
LRs shared so far: more than 200

- **File 26%**
- **Url 74%**

Also in other Conferences?

- **Community-built open repository of LRT research results**
  http://www.resourcebook.eu/shareyourlr/index.php
  - Could be the beginning of a **Language POOL**
  - Where everyone can use/**re-deposit**/create processed data of any sort
- **The LRs will go in an LREC-repository of META-SHARE**

SHARE your LRs: Resource Type distribution



http://www.resourcebook.eu/shareyourlr/index.php

LREC Open Challenge on Shared LRs

USE

- **Use the Shared LRs for whatever use/task you are interested/able to do!**
- We ask a short description on the re-use
- **We wish to promote and reward creative and original uses of available LRs**

Resource Recognition

"To Promote the LR ecosystem"

- **Facts**
  - LRs are time-consuming & costly
  - Almost **no rewards** for researchers/institutions to share, preserve, maintain, ... LRs
  - The **entire ecosystem surrounding L… should be promoted** and **sustained**

Will start soon (with ELRA & LDC catalogues, LREC shared LRs, ...)

- **Actions to be taken**
  - **Open repositories of LR research results**
  - Develop a standard **protocol** for **citation** of LRs – **ISLRN**
  - Give **greater recognition** to successful LRs & their producers: **towards a Language Resource Impact Factor**

ISLRN – International Standard LR Number

- **Unique Identifier** that allows to name and discover LRs for HLT, internet free
- ISLRN is not about access, not about rights, not about archiving
- ISLRN is not an obligation but rather a best practice
- ISLRN is not a "legal deposit“
- → ***A service to the whole community***

- **Launch the ISLRN Portal asap**
  - Promote the use of ISLRN for LR citations & as the “LR Impact factor” basis
  - Promote at LREC (e.g. shared LRs)

International Cooperation

"Promote synergies among initiatives at international level"

And communities!

- **Facts**
  - Cooperation among countries & programs essential to drive the field forward in a coordinated way, avoid duplication of efforts & fragmentation

Started NLP12

- **Actions to be taken**
  - Establish an **International Forum** to share information, discuss **future policies & priorities on a global scale**
  - **Share** the **effort** for **production** of LRs between international bodies and countries
  - Maintain a **public survey** on LRT **worldwide**

Call for cooperation

NLP12 – Paris – organised by ELRA

1. **Plan conferences and work out a common coherent planning**
2. LR identification and discovery, and promotion of best practices in LR citation in publications: announce **the establishment of the International Standard Language Resource Number (ISLRN), a Persistent Unique Identifier, to be assigned to each LR.**

Experiment **replicability**, an essential feature of scientific work, would be enhanced by such unique identifier. Set up by ELRA, LDC and AFNLP/Oriental-COCOSDA, the ISLRN Portal will provide unique identifiers using a standardised nomenclature, as a service free of charge for all Language Resource providers. It will be supervised by a steering committee composed of representatives of participating organisations and enlarged whenever necessary.

3. Encourage **LRT sharing through the use of interoperable formats** and easy-to-use schemas, in particular Creative Commons.
4. **Strengthen the bridges between various communities (e.g LT** and Humanities).

** Alliance of Digital Humanities Organizations (ADHO), Association for Computational Linguistics (ACL), Asian Federation of Natural Language Processing (AFNLP), COLING Committee (ICCL), European Data Forum, European Language Resources Association (ELRA), International Association for Machine Translation (IAMT), International Committee for the Coordination & Standardisation of Speech Databases and Assessment Techniques (COCOSDA), International Speech Communication Association (ISCA), Linguistic Data Consortium (LDC), Oriental COCOSDA, Language Resource Management Agency (RMA)*

Infra

Resource Infrastructure

- **Facts**
  - Need for facilities supporting seamless access, use, re-use, trust of data
  - Coordination among infrastructural initiatives is needed

- **Actions to be taken**
  - Build a **sustainable facility for discovering, accessing and sharing data & tools**
  - Establish **international hub(s) of resources and technologies** speech and language **services**, **– Pooling of services**, **L-Apps**

➧ **Considered important also in LOD community**

Call for cooperation

To make it more "successful"!

- It's time of **reaching the BIG community** and the **BIG data**
- And **actively involving the community** at large: many users

- So that **many** are interested in depositing & downloading & **enriching & re-depositing** LRT
- So that **Funding bodies** are involved for projects' results
- So that there is **"movement", traffic ..** In an **easy** way!

BUT it must be **organised & promoted !**

- We are a **small community** ... but we have a nb of Roadmaps, Infrastructures, Associations., Standards, similar resources/tools ... for similar purposes
- As if we could waste our efforts … instead of joining forces & compete **together** with stronger communities (in the BIG DATA arena)
- Must continue the **effort towards simplification & synergies**

Call for cooperation

Synergies instead of dispersion

Collaborative /Social

Cultural change

LRs

Services

LLOD

@ LRECs

- LRE Map (2010)
- Language Library (2012)
- MT Library (2012)
- Share your LRs! (2014)
- Challenge on Shared LRs (2014)

Services for the community

Around data & metadata

Increase Community Assets

FOR

An ecology of language

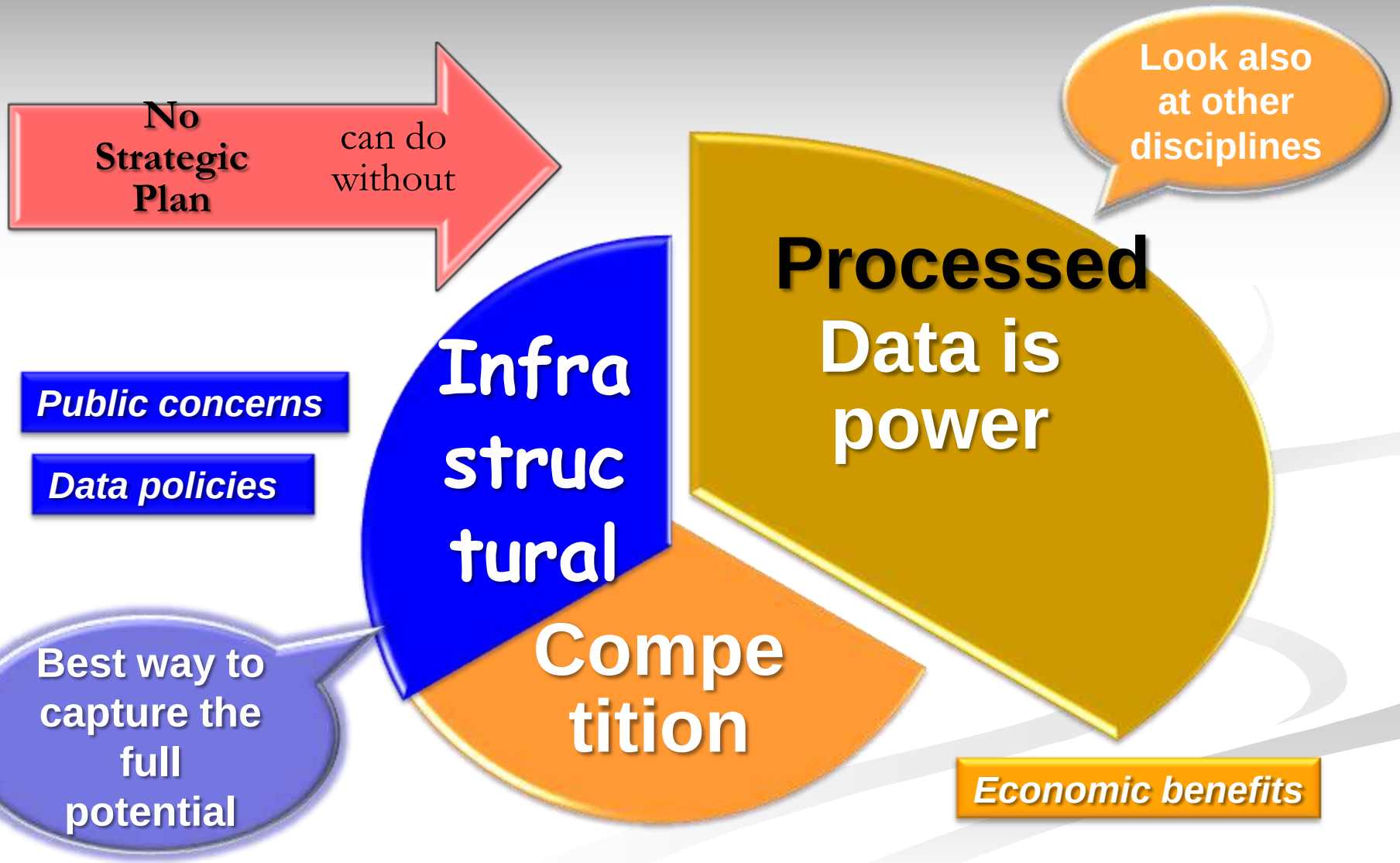
Language as a sensitive issue where we are all involved

META-SHARE

FLaReNet

BIG MORE

Recognise the Value – & solve the Conflict



BIG USE MORE

The Challenge: how to unlock the value