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
NLP is “data intensive”

Story about

Infrastructural nature
- Like railway, electricity, …

I prefer
Language Resources

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

Infrastructural issues

Here in the forefront

Also some Calls for Cooperation

N. Calzolari
FLaReNet Recommendations: a global perspective

INFRASTRUCTURE

International Cooperation

Sustainability
Recognition
Development

Documentation
Interoperability
Availability
Coverage
Resource Interoperability

“Our 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
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

**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)
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 & Metadata building as a **collaborative “shared task”**

N. Calzolari

LIDER - Athens, March 2014
Conferences & LR in the LRE Map

- LREC 2010
- COLING 2010
- ACLHLT 2011
- IJCNLP 2011
- Interspeech 2011
- LTC 2011
- RANLP 2011
- O-COCOSDA 2011
- LREC 2012
- COLING 2012
- NAACL 2013
- Interspeech 2013
- RANLP 2013

- LREC 2014
- COLING 2014
- ACL 2014
- ...

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

http://www.resourcebook.eu/
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*
**LRs as services**

- Composite access
- Web-services for Visualisation, Analysis, ...
- Extracting, Adapting, Merging, Linking, ...
- ...

Another evolution

TO

SERVICES

**LRs as services &**

Services around LRs
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

N. Calzolari
LIDER - Athens, March 2014
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

To 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%

Community-built open repository of LRT research results

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

- 40% Corpus
- 20.9% Lexicon
- 14.1% Tagger/Parser
- \(\text{Ontology} \)
- \(\text{Corpus Tool} \)
- \(\text{Annotation Tool} \)
- \(\text{Evaluation Data} \)
- \(\text{Language Resources/Technologies Infrastructure} \)
LREC Open Challenge on Shared LRs

- 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


Linked to LLOD?
“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 LRs should be promoted and sustained

**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

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

N. Calzolari

LIDER - Athens, March 2014
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”

■ Facts

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

■ 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
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)
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** for speech and language **services**, - **Pooling of services**, **L-Apps**

> Considered important also in LOD community

---

N. Calzolari
LIDER - Athens, March 2014
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!

- 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**

BUT it must be **organised & promoted**!
LRECs
- LRE Map (2010)
- Language Library (2012)
- MT Library (2012)
- Share your LRs! (2014)
- Challenge on Shared LRs (2014)

Around data & metadata

Services for the community

Increase Community Assets

An ecology of language

Language as a sensitive issue where we are all involved
Recognise the Value – & solve the Conflict

No Strategic Plan can do without

Processed Data is power

Public concerns

Data policies

Best way to capture the full potential

Economic benefits

Best way to capture the full potential

Data policies

Public concerns

Processed Data is power

Competition

Infrastructural

Big

More

Look also at other disciplines
The Challenge: how to unlock the value

- APPs around LRs
- New Services
- Exploitation
- Innovation
- LRs/Data
- Competition
- Openness
- Sharing
- Interoperability
- Legal issues
- Collaborative
- Pool ALL projects results

N. Calzolari

LIDER - Athens, March 2014