The LREMap for under-resourced languages

Riccardo Del Gratta, Francesca Frontini, Fahad Anas Khan, Claudia Soria

CNR-ILC, Italy

Joseph Mariani

LIMSI-CNRS & IMMI, France

Introduction

- European Internet users: 518,512,109 in 2012 (Source: InternetWorldStats)
- At least 21 European languages in danger of digital extinction (META-NET White Paper)
- “Where language is the very stuff of our digital system – customer interactions, employee conversations, technical and scientific knowledge, cultural and social objects of all kinds – the era of the Lingua Franca is over. Interacting across the many languages of the digital world is no longer optional.” (LT-Innovate)

Basic problems

- Lack of knowledge about existing resources; resources poorly documented or not documented at all
- Inconsistent use of common metadata elements
- Need to provide a simple and easy way to encourage documentation of resources (non-documentation = non-existence)
- Provide a monitoring instrument for surveying use and usability of LRs in the field (which resources for which applications?)

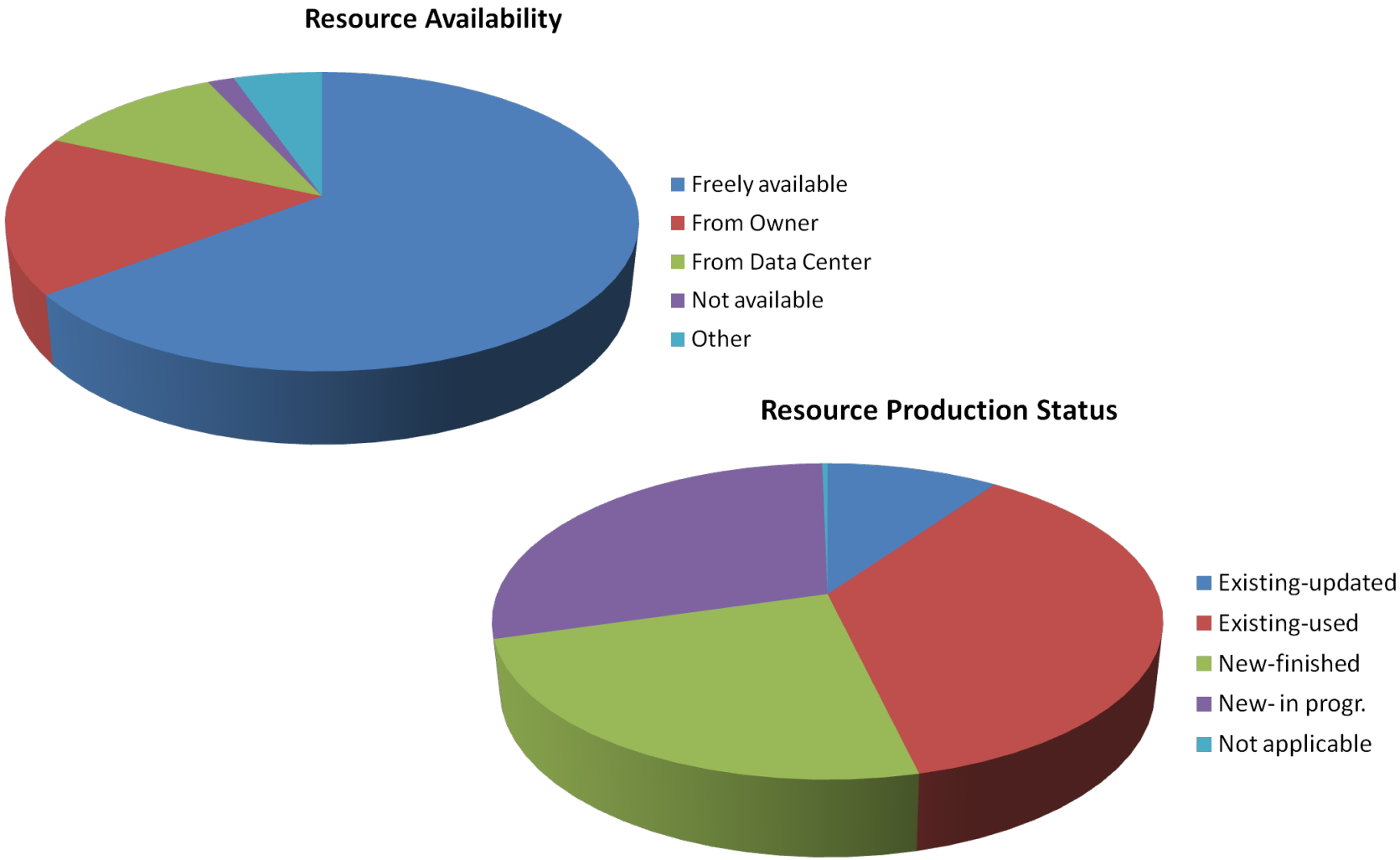
The LREMap

- Born as:
  - A FLaReNet-ELRA initiative, May 2010
  - An entirely new instrument to capture community knowledge about LRs and their uses
- Resulting in:
  - A community-based, collaboratively built meta-resource
  - A collection of uses and applications of available LRTs either developed or used for research
  - About 7000 records from 10 conferences

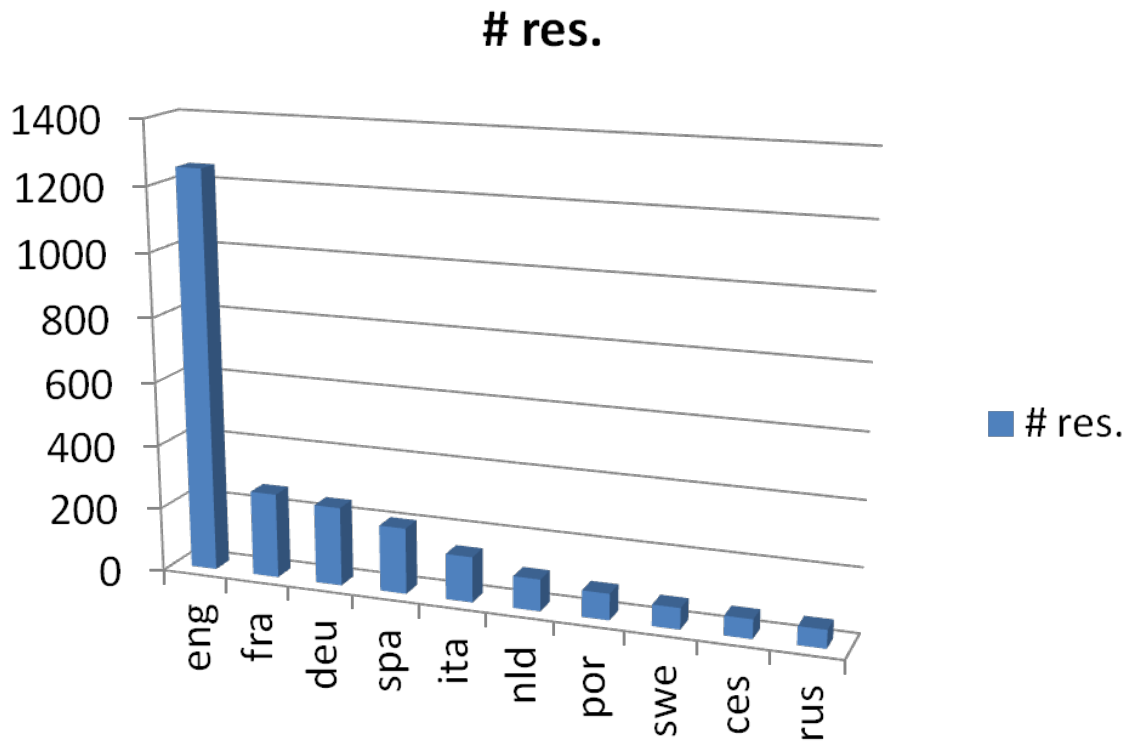
Main features

- Each resource described using a minimal metadata set (language, type, use, status, availability, etc.)
- Input gathered bottom-up, from real users and developers
- Change the way in which documentation happens
  - The community has the knowledge
- Focus on actual uses and applications in research

Type of information derivable

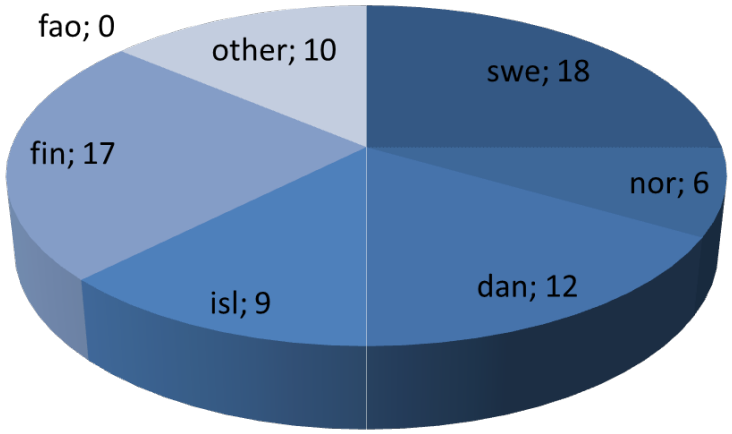


Resources per language

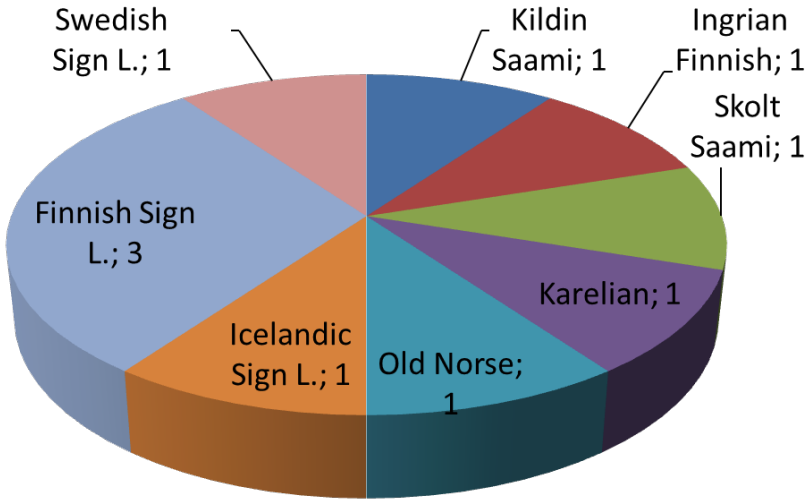


The LRE Map and less resourced languages

Scandinavian Languages



"Other": regional and heritage languages



Resources for Icelandic

Language	Type	Name	Modality
Icelandic	Corpus	Tagged Icelandic Corpus (MIM)	Written
	Corpus	MIM-GOLD	Written
	Corpus	Icelandic-Bulgarian Edda	Multimodal/Multimedia
	Corpus Tool	Crowdsourced Icelandic	Written
	corpus, lexicon and grammar/language model	SignWiki	Sign Language

Uses and application: Language Matrices

- Provide a concise picture of what is available in terms of LRs, focusing on gaps
- Offer a comparative tool that can be used for research and development

Language Matrices

		Regional and minority																									
Resource Modality	Resource Category	Aragonese	Asturian	Basque	Breton	Catalan	Corsican	Dreents	Faroese	Frisian	Galician	Limburgan	Lombard	Low German	Lule Saami	Luxembourgish	North Saami	Occitan	Romansh	Scottish Gaelic	Scots	Sicilian	South Saami	Swiss German	Venetan	Welsh	
Written	Data	2	1	18	3	23	1	1	2	2	7	1	1	1		1		1	3	1	1	1		1	2	2	76
	Evaluation			1																							1
	Guidelines			2		1																					3
	Tools	1	2	12		11			1		2				1		1						1		1	2	35
Speech	Data			6		5			1		3					1								2			18
	Evaluation					1																					1
	Tools			3					1																		4
Multimodal	Data					1																					1
	Tools			1																							1
Sign language	Data					1																					1
Written/Speech	Data			1						1																	2
Not applicable	Data			1																							1
Total		3	3	45	3	43	1	1	5	3	12	1	1	1	1	2	1	1	3	1	1	1	1	3	3	4	144

A background for BLaRK

Resource Category	Resource Type	Aragonese	Asturian	Basque	Breton	Catalan	Corsican	Dreents	Faroese	Frisian	Galician	Limburgan	Lombard	Low German	Lule Saami	Luxembourgish	North Saami	Norwegian Bokmal	Norwegian Nynorks	Occitan	Romansh	Scottish Gaelic	Scots	Sicilian	South Saami	Swiss German	Venetan	Welsh
Data	Corpus	2	1	10	2	12	1		1	2	4	1	1	1		1		1	2	1	2	1	1	1			2	1
	Lexicon			6		8		1			3										1					1		1
	Ontology																											
	Grammar/Language model					1																						
	Terminology			2		2																						
Tools	Annotation Tool			1		2																						
	Tokenizer		1			3					1																	1
	Tagger/Parser		1	4		3			1		1				1		1								1			1
	Named Entity Recognizer			1																								
	Word Sense Disambiguator			1																								
	Language Identifier																											
	Transcriber																											
	Machine Translation to	1		1																								
	Other			4		3																					1	

Why the LREMap in LLOD?

Use of important and widely accepted ontologies

- bibo for articles
- foaf for authors
- geonames for geographic features
- lexvo for languages

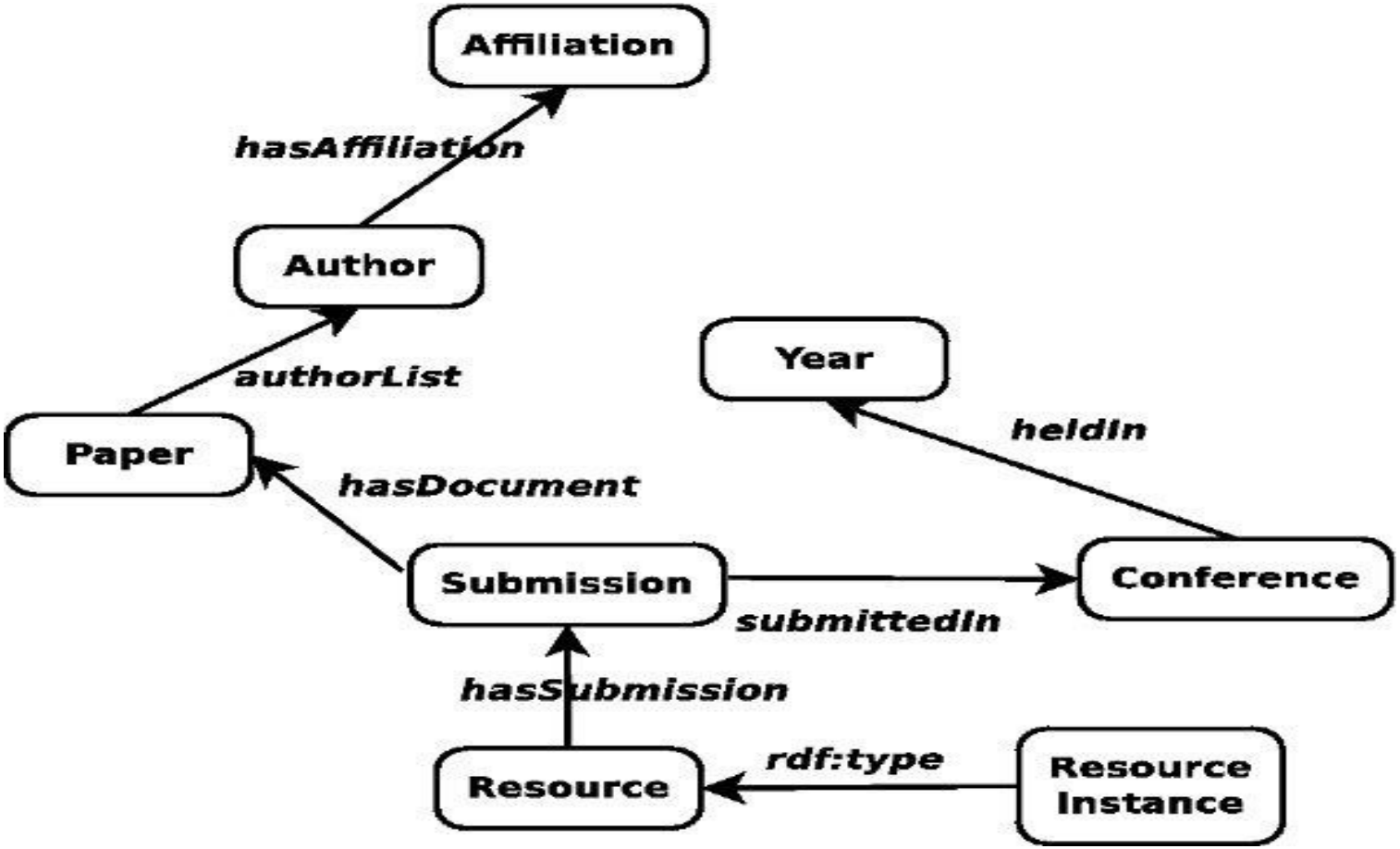
Definition of specific schemes

- language resource: the generic LR, a prototype
- language resource instance: the single resource described by authors

Data are normalized and light

- Data gathered by the LREMap are quite noisy. Names, types and languages have been normalized
- RDF serialization follows “one language resource one RDF file”: all we need to know in few Kb

LREMap scheme



Riccardo Del Gratta, Gabriella Pardelli and Sara Goggi: **The LRE Map disclosed, a roadmap to LOD**
Poster Session **P52** - Linked Data **9.45 - 11.25 DAY 3 (30/05/2014)**

Conclusions

- An instrument for…
  - Enhancing availability of information about LRs
  - Highlighting best tools and resources
  - Monitoring usability of existing resources over a range of different tools and applications