





Success stories of Collaboration in Social Sciences and Humanities

Monica Monachini

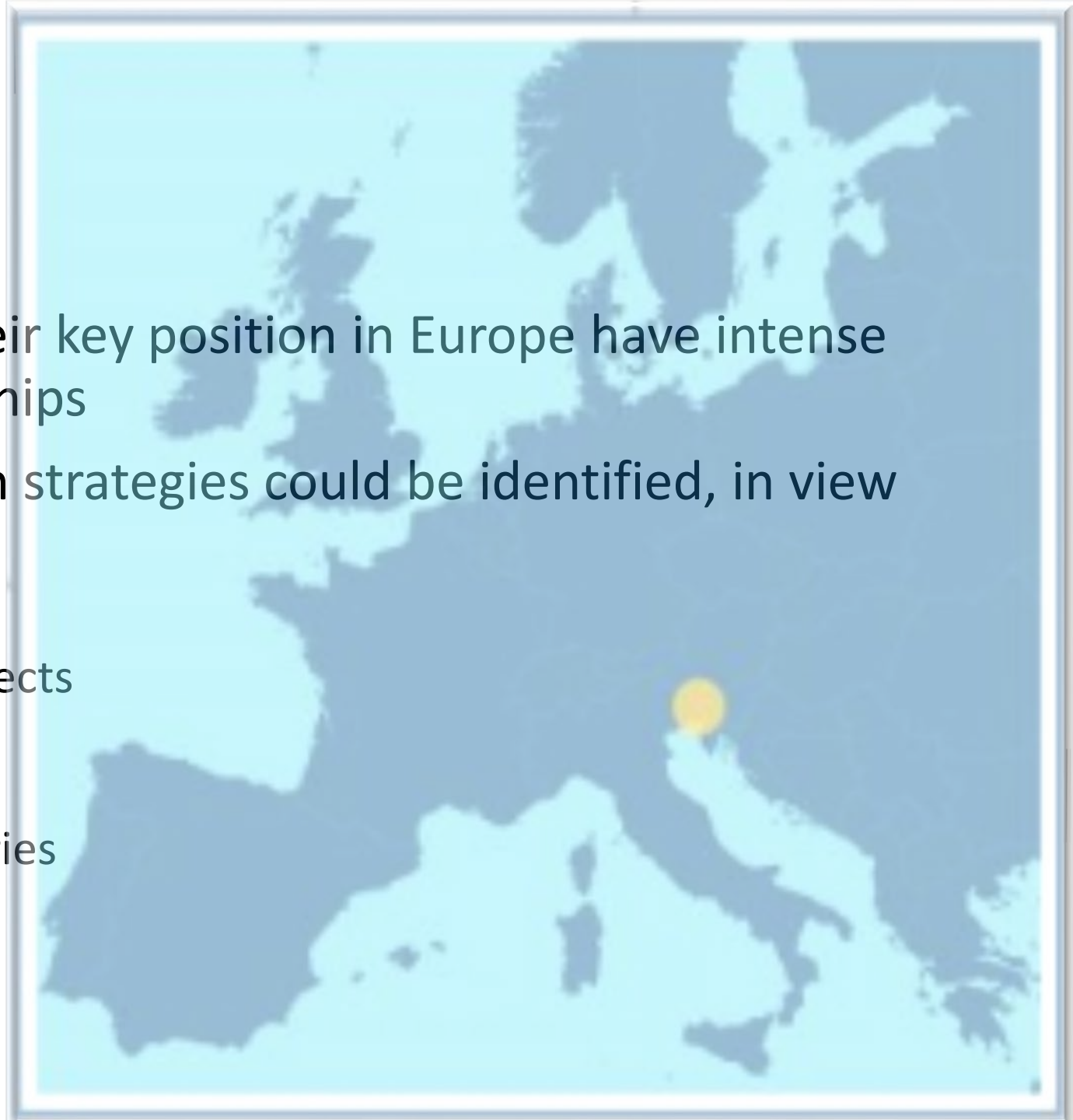
Institute of Computational Linguistics "A. Zampolli" (Pisa) – ILC

National Research Council – CNR

Dept. of Social Sciences and Humanities – DSU

Italy and Slovenia – proximity effect

- Slovenia and Italy thanks to their key position in Europe have intense scientific and cultural relationships
- commonalities and synergies in strategies could be identified, in view to
  - foster networking
  - design and implement joint projects
  - identify common strategies to tackle the major societal challenges facing the two countries



Italy and Slovenia – collaboration in research

- Example of good Slovenian-Italian collaboration, where ILC-CNR the Jožef Stefan Institute are involved
- identification of interdisciplinary themes in areas where collaborations between the two countries already exist since many years
- presentation of common activities of today
- Areas of innovation and strategic activities to tackle the major research and societal challenges

Institut
"Jožef Stefan"
Ljubljana, Slovenija

Institute for Computational Linguistics
«Antonio Zampolli»
National Research Council of Italy

...dates back to 1992

- was one of the main themes at the Ministerial meeting of the OECD (Organisation for Economic Co-operation and Development) Committee on Scientific and Technological Policy, held in Paris on 10-11 March 1992
- to coordinate scientific and technology projects between the Community and the countries of Central and Eastern Europe within the COPERNICUS (Community of Pan European Research Networks of Interdisciplinary Centres and Universities in Sciences) plan of action



A Copernicus project: Language Engineering …



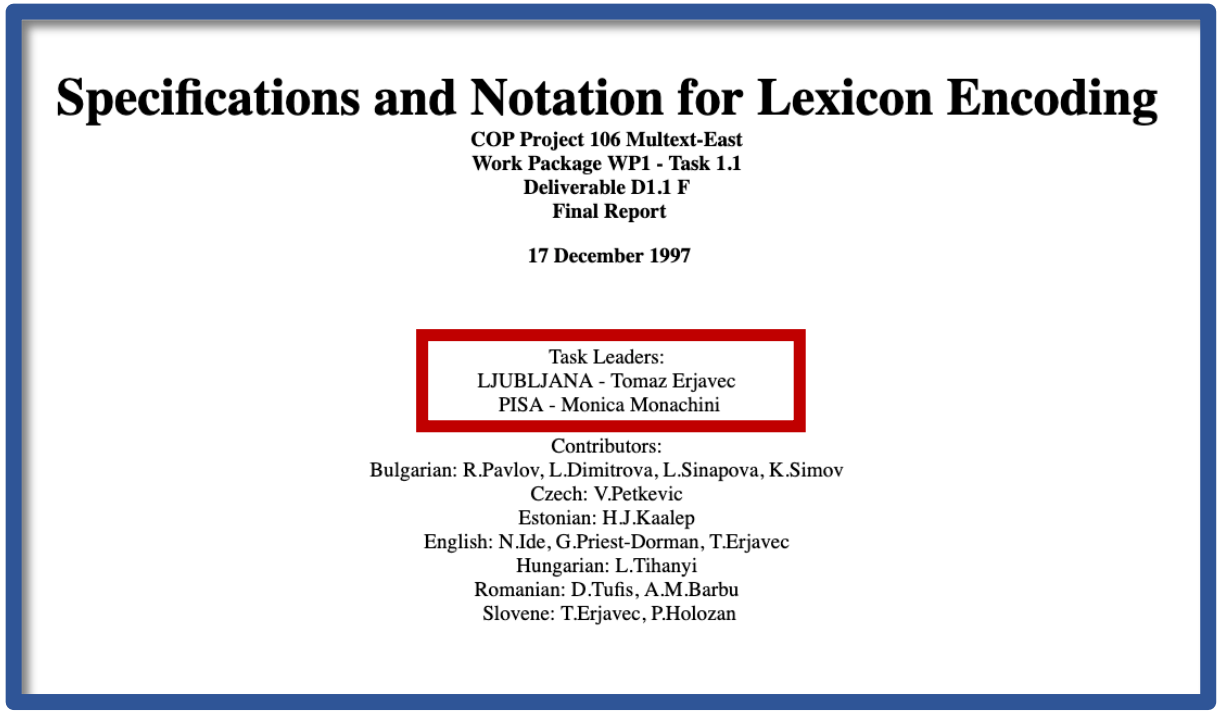
- MULTEXT, one of the largest EU projects in the domain of
  - language tools and resources
  - integration by six major European companies into NLP applications (term extraction and machine translation lexicon generation)
- MULTEXT and MULTEXT-East was a unique network of more than twenty academic research centers and companies,
  - developing and using common lingware and methodologies,
  - producing the first annotated large-scale multilingual corpus for 12 EU and CEE languages.

… in particular, Language Resources and Standards

Participants (11)

Sort alphabetically ⇕ Expand all

- Amebis D.O.O — Slovenia
- Byll Software LTD — Czechia
- Centre for Machine Learning and Natural Language (RAS) — Romania
- Charles University — Czechia
- Institut of Mathematics (BAS) — Bulgaria
- Istituto di Linguistica Computazionale — Italy
- Jozef Stefan Institute — Slovenia



Specifications and Notation for Lexicon Encoding

COP Project 106 Multext-East
Work Package WP1 - Task 1.1
Deliverable D1.1 F
Final Report

17 December 1997

Task Leaders:
LJUBLJANA - Tomaz Erjavec
PISA - Monica Monachini

Contributors:
Bulgarian: R.Pavlov, L.Dimitrova, L.Sinapova, K.Simov
Czech: V.Petkevic
Estonian: H.J.Kaalep
English: N.Ide, G.Priest-Dorman, T.Erjavec
Hungarian: L.Tihanyi
Romanian: D.Tufis, A.M.Barbu
Slovene: T.Erjavec, P.Holozan

20 years later… Language Research Infrastructures

Due to the development of the internet and other e-technologies, technological support to individual languages is extremely important



REPUBLIKA SLOVENIJA
MINISTRSTVO ZA VISOKO ŠOLSTVO,
ZNANOST IN TEHNOLOGIJO

KAZALO

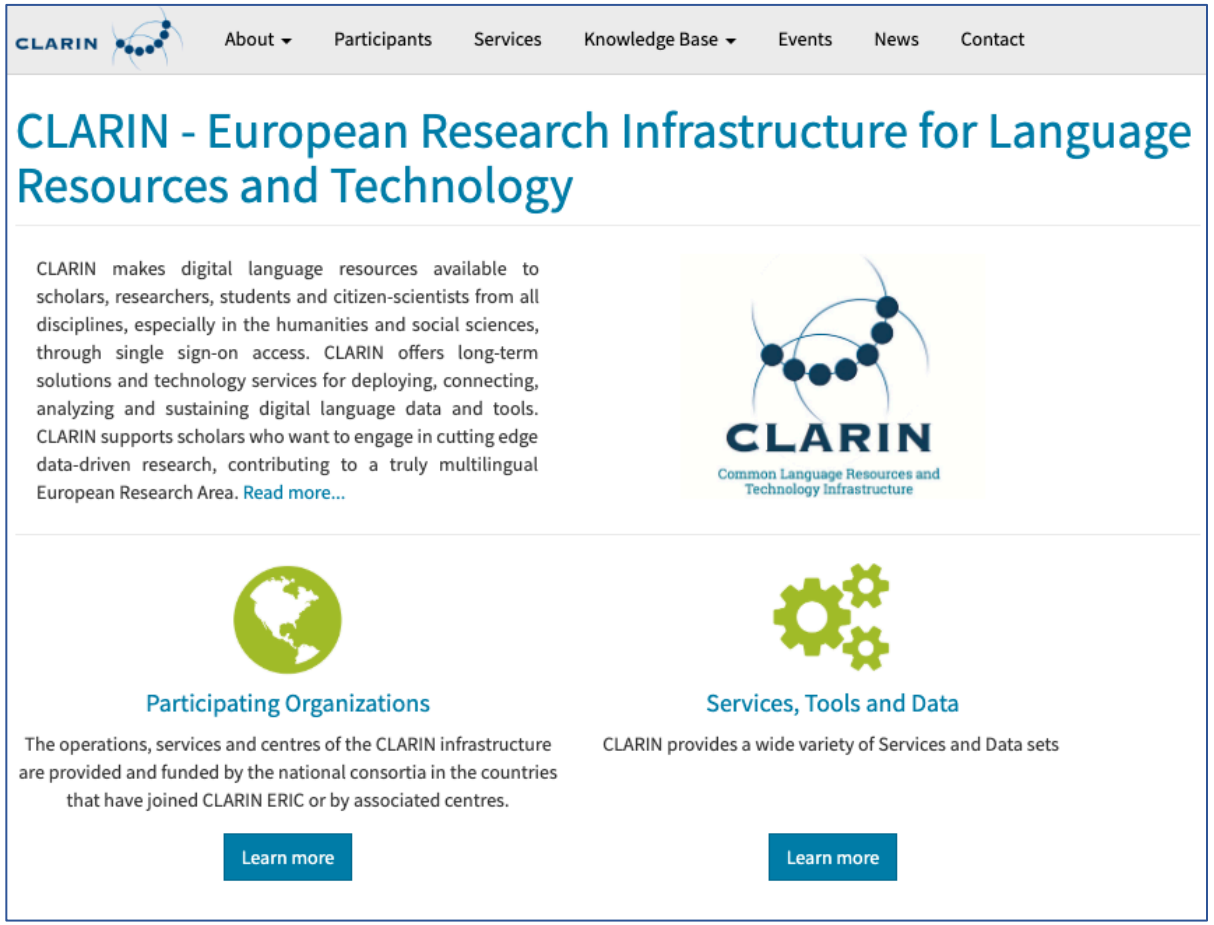
1. UVOD 3

1.1. DEFINICIJA RI 3
1.2. POMEN RAZISKOVALNIH INFRASTRUKTUR 4
1.3. AKTIVNOSTI, KI SO PRIPELJALE DO NAČRTA 5
1.4. METODOLOGIJA IZBIRE PRIORITET 6

2. OPIS PRIORITETNIH MEDNARODNIH PROJEKTOV 9

2.1. CERN 9
2.2. FAIR 11
2.3. CERIC 12
2.4. SHARE 13
2.5. ESS 15
2.6. DARIAH 17
2.7. CESSDA 19
2.8. XFEL ALI EUROFEL 21
2.9. ILL 23
2.10. BELLE II 25
2.11. LIFEWATCH 26
2.12. ELIXIR IN EATRIS 28
2.13. CLARIN 30
2.14. PRACE 31

CLARIN – for better SS and H



In 2012 CLARIN-ERIC was established with the mission to create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences.

provides easy and sustainable access to **digital language data** (in written, spoken, or multimodal form) for scholars in the **social sciences** and **humanities,**

offers **advanced tools** to discover, explore, exploit, annotate, analyse or combine such data sets, wherever they are located.

a networked federation of centers: language data repositories, service centres and knowledge centres, with single sign-on access for all members of the academic community in all participating countries.

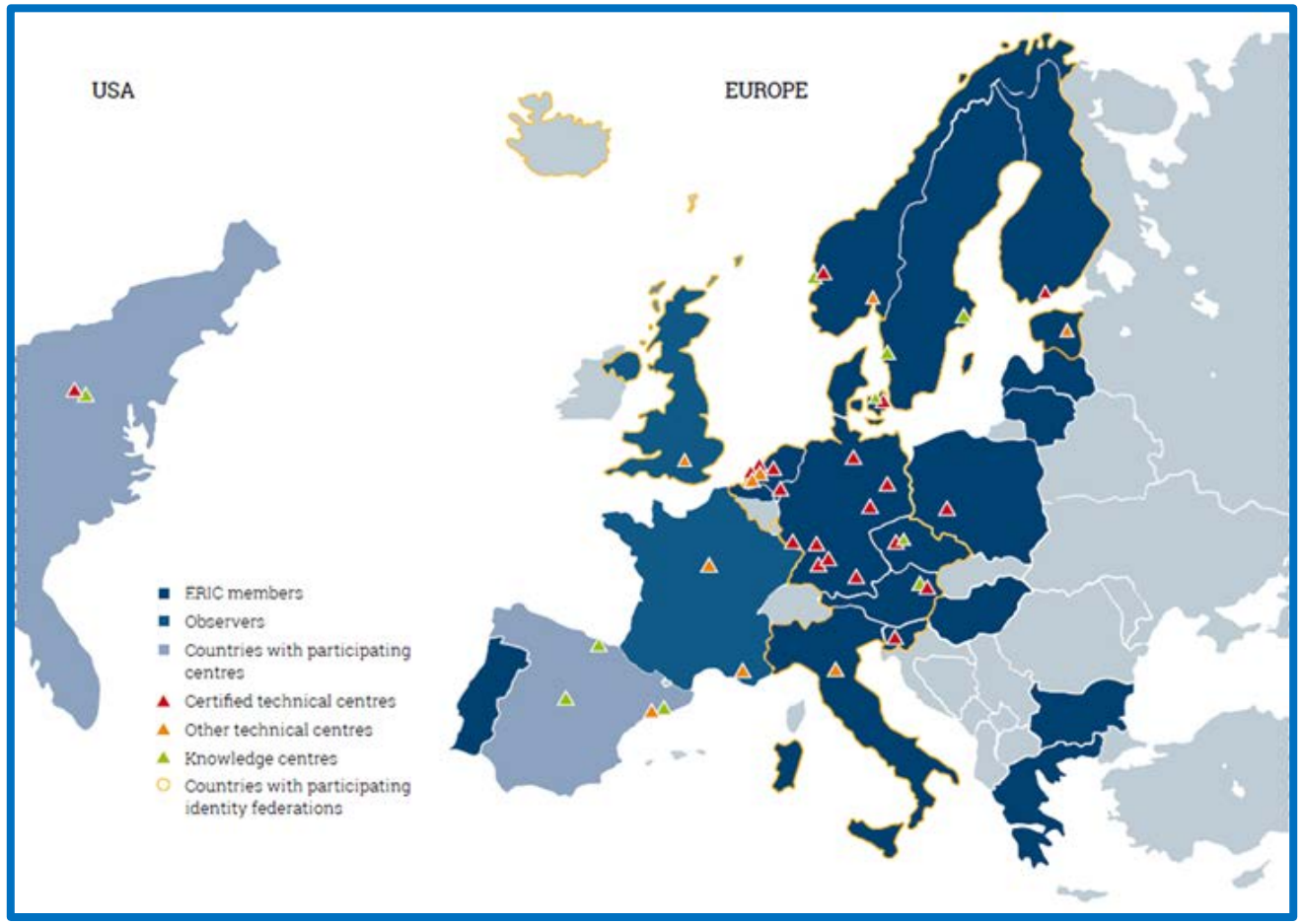
Italy and Slovenia in the CLARIN Federation

A consortium of:

1. 20 members: AT, BG, CZ, DE, DK, DLU, EE, FI, GR, HU, **IT**, LT, LV, NL, NO, PL, PT, SE, **SI**

1. 2 observers: FR, UK

1. > 40 centres



(from Darja Fisher, from Lubljana Open Science conf.)

- Slovenia was not actively involved in the preparatory phase, but had two observer institutions in CLARIN (Jožef Stefan Institute, Alpineon).
- In 2011, the Slovene Government officially approved the Research infrastructure development plan 2011-2020.
- In May 2015 Slovenia set up the CLARIN.SI language resource repository and joined CLARIN ERIC, becoming its 14th member.

- Italy was part of the ESSFRI preparatory phase 2008-2011.
- In 2015 the Italian Ministry delivers the National Research Plan 2015-2020; research infrastructures are recognized as pillar of national and international research.
- In Oct. 2015 Italy joined CLARIN ERIC, becoming its 16th member.

CLARI

Dictionaries and Lexical Resources

DICTIONARY RESOURCES

- Sloleks: Slovenian morphological lexicon

PEDAGOGICAL MANUALS

LANGUAGE

Language corpora

MONOLINGUAL

afida, 1 billion words

, 100 million words

guage Nova beseda, 318 million words

ovene IMP, 15 million words

million words

ally annotated corpus ssj500k,

manually annotated corpus jos1M, 1

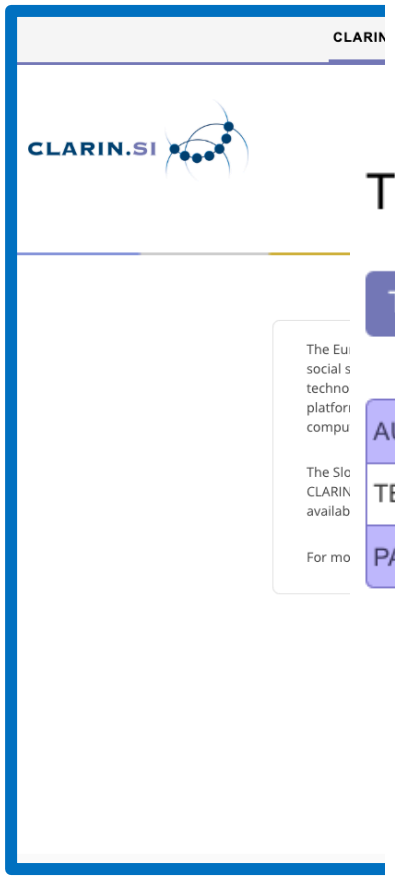
ccGigafida, 100 million words

ccKres, 10 million words

ally annotated corpus of historical

Slovene gos300k, 300.000 words

- Downloadable treebank SDT, 30.000 words



TOOLS, SERVICES, RESOURCES

Text Processing

AUTHORING TOOLS

TEXT ANNOTATION

PARSERS

- Web interface for morphosyntactic tagging and lemmatisation ToTaLe
- Web interface for morphosyntactic tagging and lemmatisation Obeliks
- Term identification and linking: Terminator

LTFW (Linguistic Tools For Weblicht)

Porting Java del tokenizzatore basato su perl sviluppato 'interno del progetto OpeNER.

ova questo strumento



ord Extractor: Merge pping Strings

re di parole multiple che un filtro che unisce sovrapposte.

uesto strumento

ord Extractor: Reorder

re di parole multiple che un filtro che riordina i dati rMW per frequenza.

uesto strumento

Organizing events, involving the scientific communities

Title	User Involvement Report for 2018
Version	1.1
Author(s)	DF
Date	06-11-2018
Status	For distribution
Distribution	BoD, NCF, SAB, GA, UI
ID	CE-2018-1284

CLARIN

Contents

Introduction **1**

1 Strategic activities **2**

1.1 Surveys of CLARIN Resource Families 2

1.2 Improving visibility of CLARIN Key Families of Resources 4

1.3 Hackathon for key Text Mining and Visualization tool developers 4

1.4 Field activities 4

1.4.1 UI events organized by CLARIN ERIC 4

1.4.2 UI events (co)funded by CLARIN ERIC 5

1.4.3 UI events endorsed by CLARIN ERIC 6

1.4.4 Contributions at third-party events 6

1.5 Overview of national CLARIN UI activities 7

1.6 Involvement in COST Actions 8

2 Supporting activities **8**

2.1 Tour de CLARIN 8

2.2 VideoLectures 9

CLARIN Annual Conference 2019

Monday, 30 September, 2019 - 16:00 to Wednesday, 2 October, 2019 - 16:00



ParlaFormat Workshop

Thursday, 23 May, 2019 - 13:00 to Friday, 24 May, 2019 - 13:00

A workshop on a proposed standard format for parliamentary data is organised by the CLARIN Interoperability Committee in close collaboration with Tomaž Erjavec and Andrej Pančur from CLARIN Slovenia.

1. Workshop goal

The goal of the workshop is to propose an outline of a standard format for parliamentary data to the research community, to assess the support for it and to identify potential or real problems for its development and wide adoption. It is intended as a preparation for the work on the CLARIN endorsed proposal for a standard encoding of parliamentary data. If the workshop shows sufficient support for the proposed standard format, it may be followed in the future by a shared task and accompanying workshop.
The standard format will be proposed, elaborated and documented by Tomaž Erjavec and Andrej Pančur. It is based on their treatment of the Slovenian parliamentary data as reported on at the ParlaCLARIN workshop at LREC, Japan in May 2018. Erjavec & Pančur used the TEI/Drama and TEI/Speech formats and created a converter from the former to the latter. The encoded corpus is available from CLARIN.SI, and will serve as the basis for the proposed format. Tomaž and Andrej will draft and present the overall architecture of the standard and its ecosystem, and carry out the subsequent work in fleshing out the draft into a well-documented and exemplified proposal that can be easily used and further developed by other researchers.

The workshop participants will be invited to comment on their proposal from the perspective of the parliamentary data they work with and their research goals, so that we get a good picture of the support for this standard, and for any real or potential problems that would lead to problems in its use and adoption.

2. Link with CLARIN's strategic priorities

The topic of the proposed workshop is linked to CLARIN's strategic priority to increase interoperability among datasets and tools, as well as to the strategic priority of creating more visibility for specific families of resources, and CLARIN's User Involvement work plan for 2018 and 2019. The

New challenges: Facing Open Science

Not stand-alone, but a component in the ecosystem of infrastructures that share the goal of making Open Science a reality

Interoperates with other e-infrastructures

Reciprocal influencing and cross-fertilisation

Interconnect researchers across national and discipline borders by offering seamless access to data and services

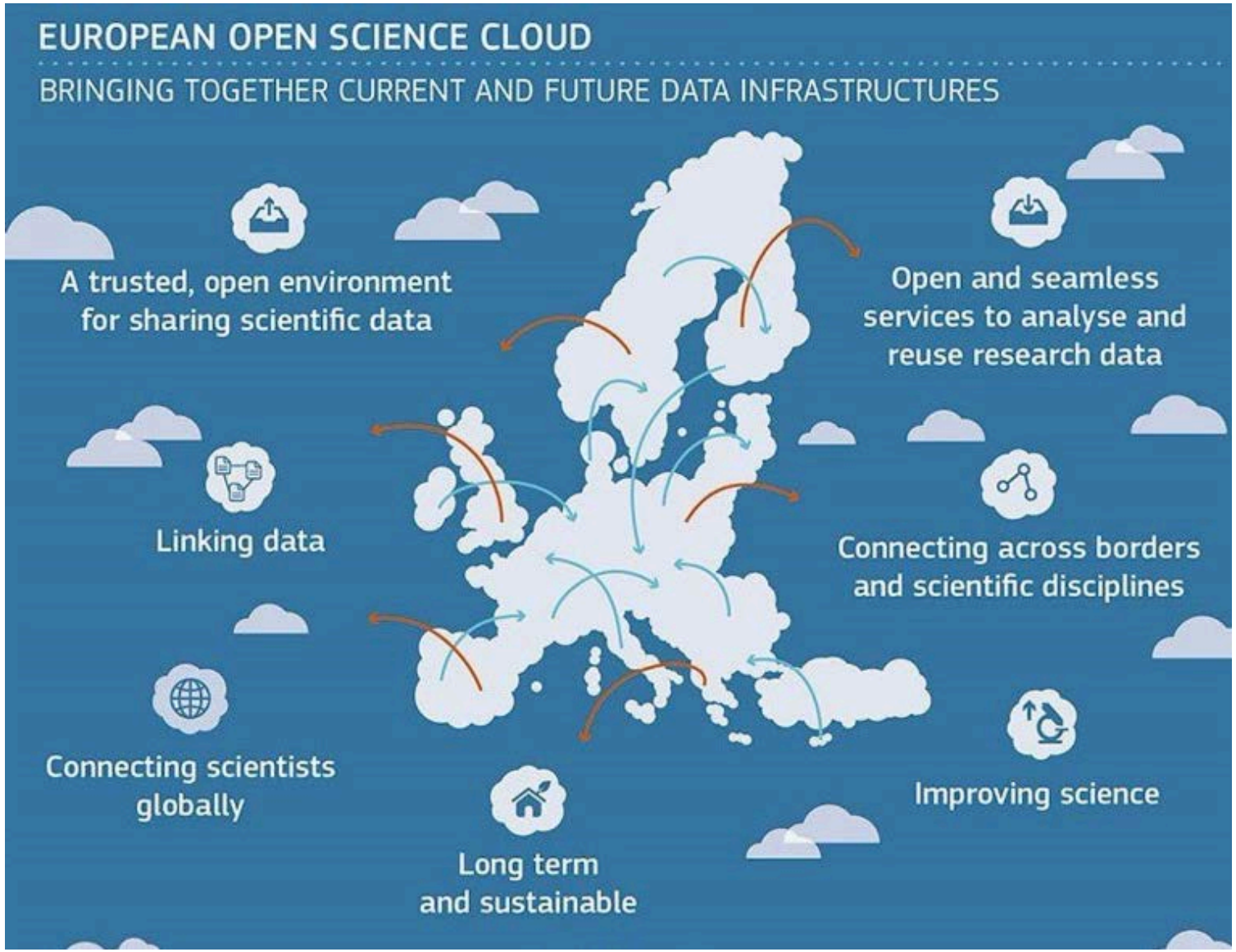


New challenges: Facing FAIR data



- persistent identifiers (PIDs), and metadatametadata are open and are harvested from centres into catalogues (including the VLO) which make them findable
- standardized communication protocol, with the option for easy authentication and authorisation when needed. CLARIN relies on the HTTP protocol for federated single sign-on. Resources are as open as possible, with restrictions when necessary
- CLARIN relies on the CMDI framework as common metadata language, including links to standardized OpenSKOS vocabularies
- re-usable data requires clear license and provenance information and adherence to community standards. CLARIN has clear recommendations on license disclosing and user-friendly ways of categorizing these

An even greater challenge: EOSC



An even greater challenge: EOSC



To promote Open Science and contribute to advancement in research in Europe there is the need to

prompt the EOS vision in practice
turn FAIR into reality
and effectively interlink people data services publications organizations



The European Union has launched the European Open Science Cloud as a supporting landscape to foster open science and open innovation: a network of organisations and infrastructures from various countries and communities that supports the open creation and dissemination of knowledge and scientific data



The creation of EOSC is aimed at removing technical, policy and human barriers, leading to knowledge creation and economic prosperity in Europe.

The European Open Science Cloud is intended to set off the ground by federating existing scientific data infrastructures that are now spread across disciplines and EU member states. This will make access to scientific data easier and more efficient



The EOSCpilot project is supporting the first phase in the development of the EOSC. The project brings together stakeholders from research infrastructures and e-Infrastructure providers and will engage with funders and policy makers to propose and trial EOSC's governance framework.



The EOSC-hub project' Jan 2018 aims to create the integration and service management structure of the European Open Science Cloud enabling an open access to research resources from a myriad of scientific disciplines via a digital hub. The hub will act as an entry point for researchers and innovators to discover, access, and use a variety of advanced data-driven resources.

Social Sciences & Humanities in the EOSC

INFRAEOSC-04-2018 – EU Cluster projects – Connecting ESFRI infrastructures through Clusters

SSHOC, one of the five European Union H2020 Programme “INFRA-EOSC-2018” recently funded cluster projects (together with ENVRI-FAIR, PANOSC, ESCAPE, EOSC-LIFE) that will leverage and interconnect existing and new infrastructures from the SSH ERICs and foster interdisciplinary research and collaboration.

19 beneficiaries, 42 organizations, 14.5ME, 1759 PMs, Duration 40 months started 1-1-2019

Main consortium partners: SSH ERICS: CESSDA (coordinator), **CLARIN**, DARIAH, SHARE, ESS, the E-RIHS network and their partner organizations

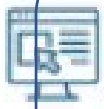
Main impacts from SSHOC



The Social Sciences and Humanities are seamlessly integrated in the European Open Science Cloud



Availability of an EU wide, easy to use SSH Open Marketplace, where tools and data are openly available



EU wide availability of high quality "cloud ready" SSH tools and high quality SSH data



EU wide availability of trusted and secure access mechanisms for SSH data, conforming to EU legal requirements



State of the art advanced through dedicated SSH data pilots cluster projects



Data sharing is accepted practice (the "new normal") among the different SSH communities

ELEXIS for Advanced Lexicography



Total cost:
4,999,967.50 €





Coordinator: **Jožef Stefan Institute, Ljubljana, Slovenia, S. Krek**



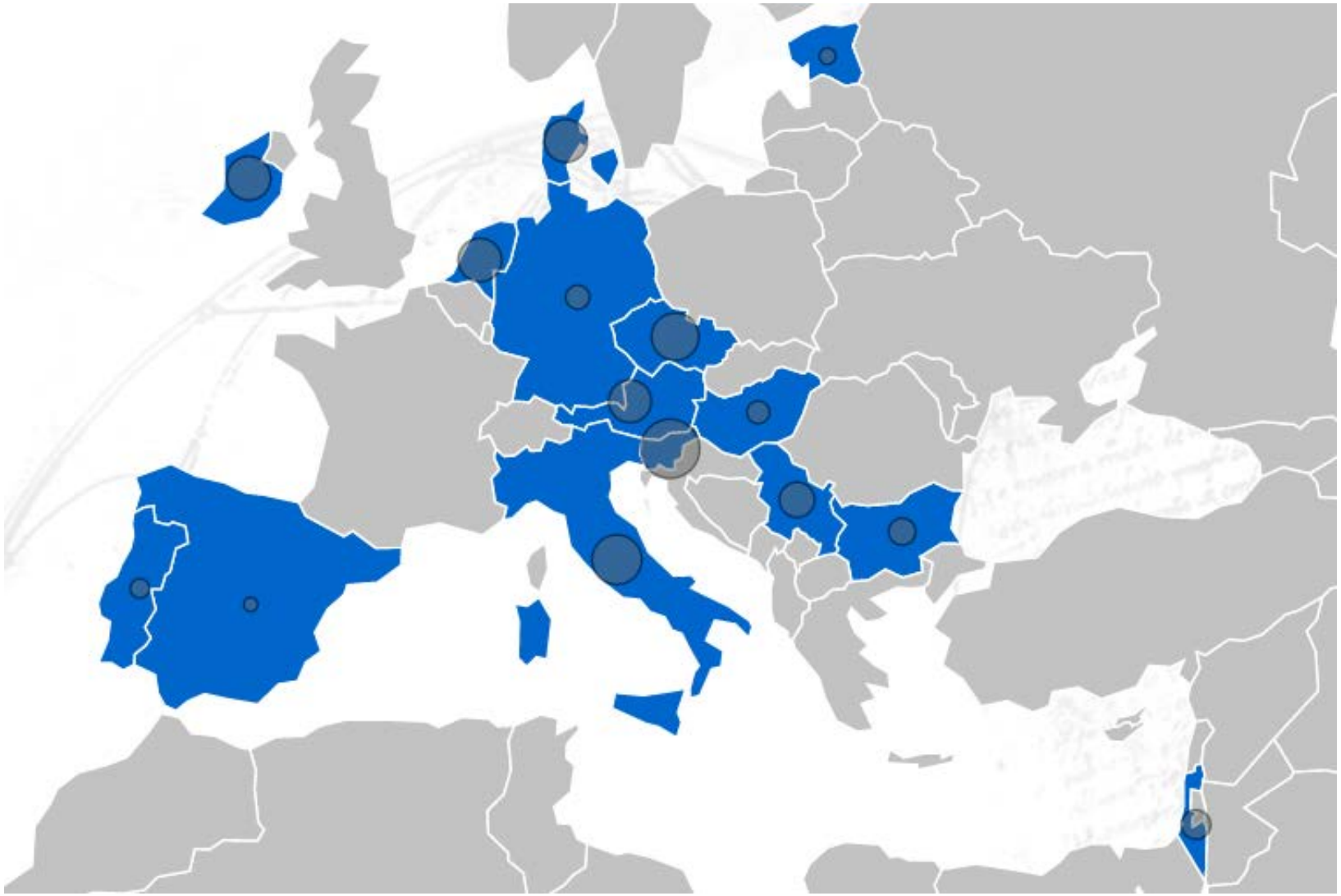
48 months (Feb 2018 Jan 2022)

Number of partners: 17 from 15 countriesWeb site: www.elex.is

Call & Topic:
INFRAIA-02-2017
Integrating Activities for **Starting Communities**



Expertise of partners



lexicographic expertise

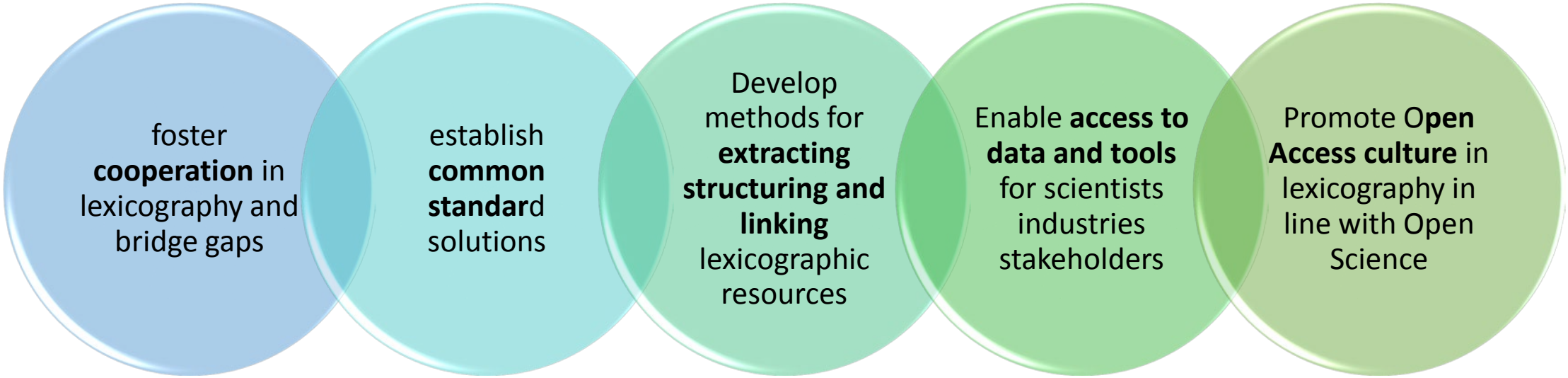
computational linguistics data expertise

expertise in standardisation

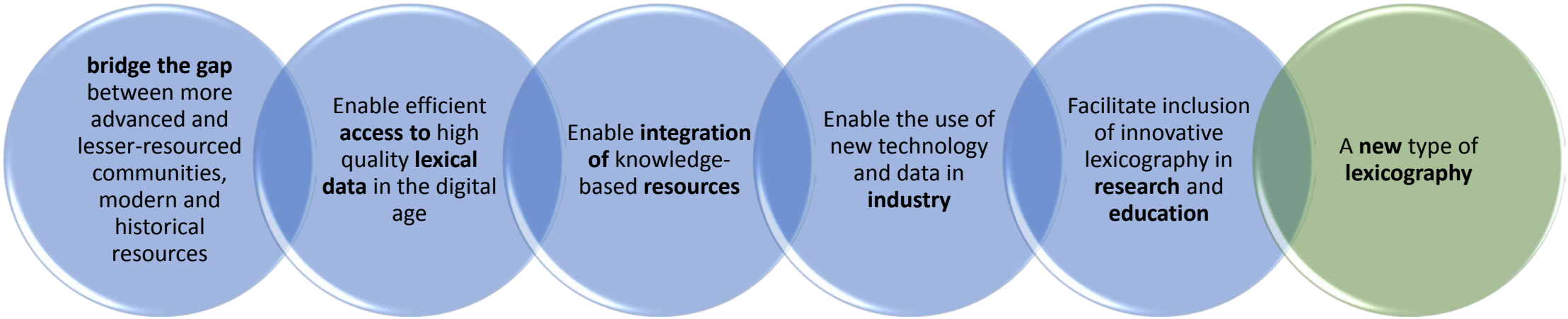
digital humanities

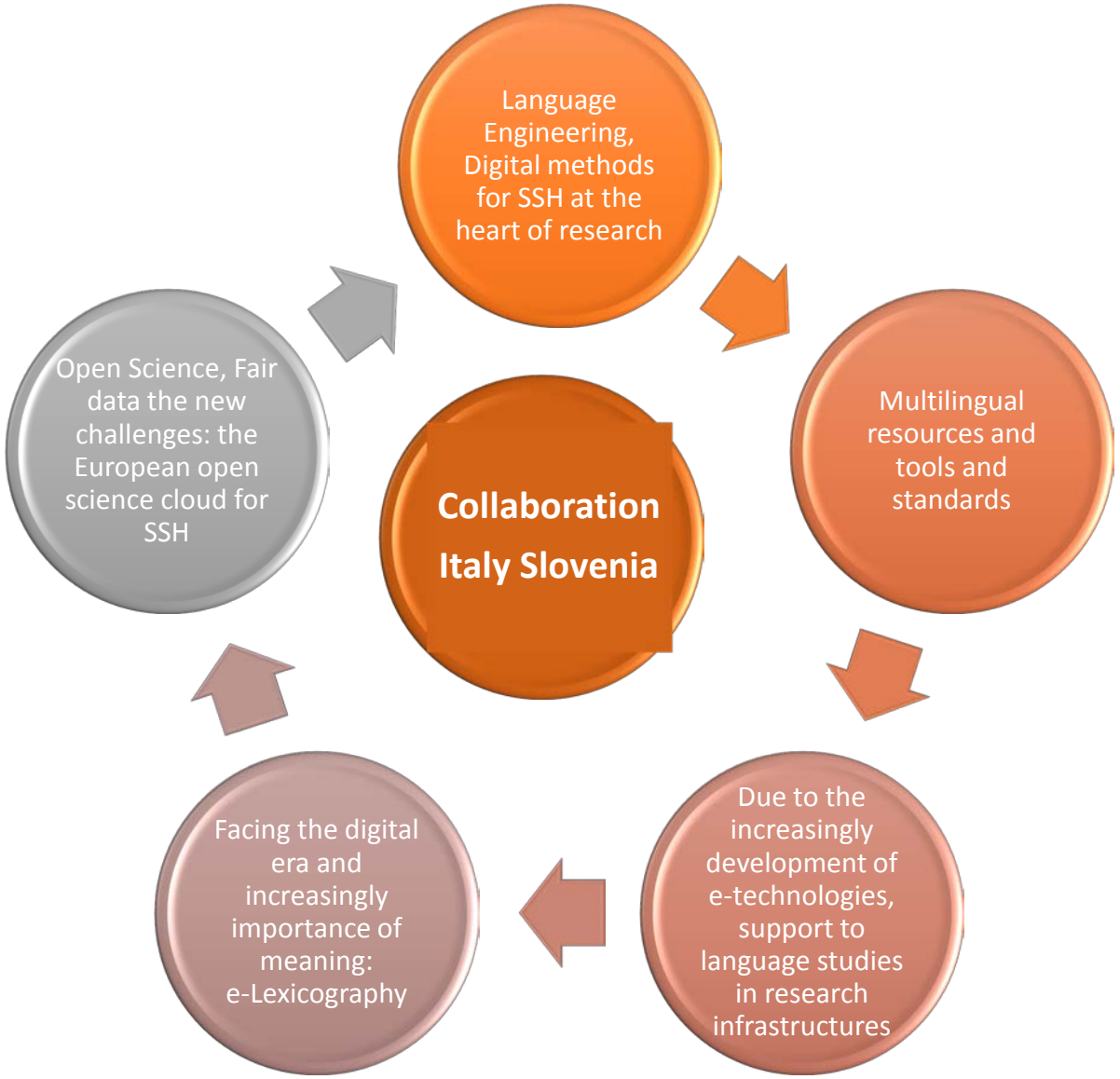
technology partners

Objectives


foster **cooperation** in lexicography and bridge gaps
establish **common standar**d solutions
Develop methods for **extracting structuring and linking** lexicographic resources
Enable **access to data and tools** for scientists industries stakeholders
Promote O**pen Access culture** in lexicography in line with Open Science

Ambitions and Impacts for the community




Language Engineering, Digital methods for SSH at the heart of research
Multilingual resources and tools and standards
Due to the increasingly development of e-technologies, support to language studies in research infrastructures
Facing the digital era and increasingly importance of meaning: e-Lexicography
Open Science, Fair data the new challenges: the European open science cloud for SSH
Collaboration Italy Slovenia