FAIR Language Resources in NLP: Stewardship, Reuse and Long-Term Sustainability

Dr Milena Dobreva
IMI-BAS and University of Strathclyde
milena.dobreva@strath.ac.uk

Dr Ivan Lambov
IMI-BAS
ilambov@math.bas.bg

Abstract

Language resources are central to NLP research, yet their long-term sustainability and structured reuse remain largely unresolved challenges. While FAIR principles are widely endorsed across research policy, their domain-specific implementation for corpora, lexicons, and evolving NLP datasets has not been systematically addressed. Current practices often prioritise release over lifecycle stewardship, leaving critical questions of versioning, governance, and measurable reuse insufficiently explored. This workshop introduces an innovative shift in perspective: from FAIR compliance to active stewardship and infrastructure design. The workshop will examine how FAIR principles can be operationalised in ways that support reproducibility, interoperability, and sustained reuse in computational linguistics. What makes this event distinctive is its integration of technical, infrastructural, and governance perspectives focusing on the NLP domain. It brings together NLP researchers and practitioners, infrastructure providers, data stewards and policy makers to address challenges that cut across metadata standards, annotation transparency, licensing constraints, and sustainability beyond project funding. Particular emphasis will be placed on mechanisms for documenting and measuring reuse. Through invited talks, peer-reviewed contributions, and an interactive FAIR assessment exercise, participants will collaboratively develop practical recommendations for sustainable language resource ecosystems.

Keywords: data stewardship, FAIR, language resources, open science, ORRI

1 Workshop Topic and Content

Language resources are foundational to computational linguistics and NLP. Corpora, lexicons, annotated datasets, benchmarks, and pretrained models grow in numbers and size. Yet while publication and repository deposit have become more common, long-term stewardship, structured reuse, and sustainability remain uneven and insufficiently addressed.

The FAIR principles (Findable, Accessible, Interoperable, Reusable) provide an important framework for research data management. However, applying FAIR to language resources raises domain-specific challenges that extend beyond repository-level compliance. Initiatives such as WorldFAIR have advanced cross-domain FAIR implementation frameworks, but language resources and NLP infrastructures have not been examined as a dedicated domain (WorldFAIR, n.d.). As a result, key questions remain underexplored: How should layered linguistic annotations be documented for long-term interoperability – and support reuse? How should evolving versions of corpora and models be stewarded? How can reuse be made visible and measurable? What governance structures ensure sustainability beyond project funding cycles?

Language data differ from many other research data types in important ways: they are iterative, socially embedded, legally constrained, and frequently reused in computational workflows. FAIR compliance in this context requires domain-sensitive implementation strategies.

This workshop aims to address this gap by examining how FAIR principles can be meaningfully applied to language resources and how Open Science practices can support sustainable language infrastructures in computational linguistics and NLP.

In this workshop, we use *stewardship* to mean the active, long-term responsibility for maintaining, documenting, versioning, and enabling reuse of language resources across research cycles. The workshop will focus on:

- Practical FAIR implementation for corpora, lexicons, and NLP datasets
- Metadata and paradata standards for linguistic data
- Documentation and versioning of evolving resources
- Licensing and ethical constraints in open language data
- Tracking and measuring reuse of language resources
- Governance models for sustainable language infrastructures

By bringing together researchers, infrastructure providers, and data stewards, the workshop aims to move from short-term dissemination toward long-term sustainability of language resources in NLP.

2 A list of invited speakers, if applicable, with an indication of which ones have already agreed

We plan to invite **Dr Francesca Frontini** from ILC-Pisa, Italy and CLARIN to introduce the topic of FAIR issues in language resources, **Andrew Clark**, a data steward, Linguistic Research Infrastructure (LiRI) – Switzerland to talk about the challenges of data stewardship for language resources, and **Dr Beth Knazook** from the Digital Repository of Ireland, who led the work on the cultural heritage case study in WorldFAIR project and could introduce the challenges in defining what FAIR data means in another digital humanities-related domain (digital heritage). We have not invited them yet.

3 An estimate of the number of attendees: 25-40

4 A discussion of measures planned to ensure the workshop is successful and productive for both in-person and virtual attendees.

Structured Programme

- Clearly defined thematic sessions

We anticipate a day-long event, with the three speakers introducing the topics of positioning FAIR data for language resources (Francesca), stewardship (Andrew) and challenges (Beth). Each topic will have the invited speaker's introduction in 30 mins and up to 3 short papers of 12 minutes, followed by discussion.

We will wrap-up the event with a moderated panel discussion on lessons learned and next steps.

- Interactive Engagement
  - Pre-workshop survey to identify participants' priorities
  - Breakout session
  - Reporting and synthesis session to identify shared recommendations
- Post-Workshop Outputs
  - Open access publication of slides and materials
  - Summary report synthesising recommendations
  - Exploration of a follow-up working group on FAIR language infrastructures with CLARIN

We are open to discussing hybrid participation if the conference supports it.

We will also use the channels of the pilot programme on open science skills to spread the event widely and connect to the activities on competence building in data stewardship in Bulgaria.

5 A description of any shared tasks associated with the workshop and an estimate of the number of participants.

We propose a voluntary interactive component entitled: **"FAIR Assessment Exercise for Language Resources"**

Participants will work in small groups to evaluate a short case study describing a published language dataset. They will assess:

- Metadata completeness and quality
- Interoperability readiness
- Licensing clarity
- Documentation of annotation processes
- Sustainability provisions

- Reuse potential

We aim to involve all participants.

The expected outcomes include:

- A community-developed checklist for FAIR and sustainable language resources
- Draft recommendations for repository and infrastructure design

6 A description of special requirements and technical needs.

If the venue allows a hybrid event, we would be happy to explore the practicalities and open the workshop to online participants as well. However, we plan for all speakers to be taking part face-to-face. The shared task can also be moderated with the online participants forming a separate group and feeding back.

7 If the workshop has been held before, a summary of where it was held, the number of submissions and acceptances, and the number of attendees.

This is our first attempt to tackle the topic.

8 The names, affiliations, and email addresses of the organisers, with a brief statement on their research interests, expertise, and organizational experience.

Assoc. Prof. Milena Petrova Dobreva
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (IMI-BAS), Bulgaria, ERA Chair, FOCUS – Fostering Digital Cultural Heritage via Open Innovation and Open Science
University of Strathclyde, United Kingdom
Email: milena.dobreva@strath.ac.uk

Assoc. Prof. Milena Dobreva is an established researcher in Open Science, digital research infrastructures, and sustainable data practices in the arts, humanities, and social sciences. Her work focuses on research data stewardship, FAIR implementation, digital cultural heritage, and infrastructure governance. She has coordinated and contributed to European initiatives on Open Science capacity building and research infrastructure development, and has extensive experience organising international academic workshops and conferences related to digital innovation and data sustainability.

Dr. Ivan Lambov
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (IMI-BAS)
Email: ilambov@math.bas.bg

Dr Ivan Lambov is an Assistant Professor at IMI-BAS, where he works at the intersection of AI, information technologies, and cybersecurity. His research interests include AI methods, cyber-physical systems, de-centralised technologies such as blockchain, and their application in secure and intelligent information systems. He contributes to research on integrating advanced computational methods with robust data governance, aligning closely with themes of sustainability and responsible Open Science.

9 A list of Program Committee members, with an indication of which members have already agreed.

We have not contacted potential members of the PC but are planning to invite the following academics:

Prof. Olivier Baude
CNRS / Huma-Num, France
Expertise: National research infrastructures, preservation and dissemination of linguistic resources.

Prof. Philipp Cimiano
Bielefeld University, Germany
Expertise: Linguistic Linked Open Data, semantic interoperability, structured representation of language resources.

Dr Edward Gray
DARIAH-ERIC
Expertise: Research infrastructures, interoperability frameworks, digital SSH infrastructures.

Dr Ruslana Margova
GATE Institute
Expertise: Computational linguistics, NLP for less-resourced languages, language of disinformation

Prof. Petya Osenova
Sofia University “St. Kliment Ohridski”, Bulgaria

Expertise: Computational linguistics, NLP for less-resourced languages, language resource development and evaluation.

Prof. Christophe Parisse

Université Paris Nanterre & CNRS (Ortolang / CORLI), France

Expertise: Corpus linguistics, TEI encoding, language resource formats, sustainability of linguistic corpora.

Prof. Eva Soroli

Université de Lille, UMR STL (Savoirs, Textes, Langage), France

Expertise: FAIR principles in linguistics, digital infrastructures for language data, Open Science in SSH.

Prof. Matthias Urban

University of Tübingen & CNRS (DDL – Dynamique du Langage), Germany/France

Expertise: Cross-linguistic data formats (CLDF), FAIR data modelling, linguistic databases.

Addition: Draft Agenda

09:00–09:15 – Welcome and Introduction

Opening remarks by the organisers. Framing the workshop: stewardship, reuse, and long-term sustainability of FAIR language resources in NLP.

Session 1: FAIR Implementation in Language Infrastructures

09:15–09:45 – Keynote: Dr Francesca Frontini (ILC-Pisa, Italy; CLARIN)

FAIR Implementation Challenges for Language Resources

09:45–10:45 – Short Paper Presentations

(3–4 peer-reviewed papers, followed by 10 minutes moderated discussion.)

10:45–11:00 – Coffee Break

Session 2: Stewardship and Lifecycle Management

11:00–11:30 – Keynote: Andrew Clark (Linguistic Research Infrastructure – LiRI, Switzerland)

From Resource Release to Lifecycle Stewardship

11:30–12:30 – Short Paper Presentations

12:30–13:30 – Lunch Break

Session 3: Cross-Domain Perspectives on FAIR

13:30–14:00 – Keynote: Dr Beth Knazook (Digital Repository of Ireland; WorldFAIR Cultural Heritage Case Study Lead)

Interpreting FAIR Across Domains: Lessons for Language Resources

14:00–15:00 – Short Paper Presentations

15:00–15:15 – Coffee Break

15:15–16:00 – FAIR and Stewardship Assessment Exercise

Breakout groups evaluating a case-study language resource using structured FAIR and sustainability criteria.

16:00–16:45 – Panel Discussion

Selected contributors, including policy makers, academics, and data stewards, reflect on sustainability challenges, governance models, and future directions.

16:45–17:00 Closing Remarks and Next Steps

Note: *we have not confirmed keynotes, but in case of acceptance, they will be our first choice. We have additional options if anyone is not available. We also could support the expenses for bringing the keynote speakers.*

Timetable for Workshop upon acceptance (coordinated with the conference timeline)

- Inviting PC and invited speakers in the week after acceptance
- Issuing CFP: 15 March 2026
- Deadline to submit short papers (4-6 pages): 22 April 2026
- Authors' notification: 22 May 2026
- Event: 7 September 2026

Acknowledgments

We would like to acknowledge the support of FOCUS ERA Chair - Fostering Digital Cultural Heritage via Open Innovation and Open Science, and of the Pilot Programme for Capacity Building of Higher Education Institutions and Research Organisations in the Creation and Use of Research Data through the Principles of Open, Responsible and Citizen Science.

References

WorldFAIR. n.d. Project website, WordlFAIR. https://worldfair-project.eu/

ADDENDUM 06 March 2026

Will the proposers rely on the CLIB organisers to provide information about the workshop on the CLIB website, or do they intend to maintain their own website for the workshop? If the former, please provide the exact information you would like to appear on the CLIB web page. If the latter, please provide a link to the workshop website.

We will maintain our website. For the time being we have a draft of the CFP and we will add a page on the FOCUS website. The CFP is on https://easychair.org/cfp/FAIR-CLIB2026 - next week we hope to complete most information on keynote speakers and PC. So far Beth Knazook confirmed; Francesca Frontini is unavailable but we expect her recommendation for a potential speaker.

Below is a tile we plan to use for social media publicity when the sites are properly updated. We are also planning to produce tiles with the speakers info which we will share on social media.

FOCUS
Workshop organiser
7–9 Sept 2026
7th International Conference Computational Linguistics in Bulgaria
Sofia, Bulgaria
(CLIB 2026)
FOCUS co-organises a pre-conference workshop on FAIR Language Resources in NLP:
Stewardship, Reuse and Long-Term Sustainability
CliB

Will the proposers rely on the CLIB organisers to set up an EasyChair platform for workshop submissions? Please note that the workshop organisers should manage their own campaign, call for papers, reviewing, and acceptance of workshop papers.

NO, we already configured our submission page https://easychair.org/my2/conference?conf=fairclib2026

Will the proposers rely on the CLIB organisers to publish the workshop proceedings, and if so, what kind of assistance do they expect?

Yes it will be great to have our proceedings as a supplement to CLIB proceedings – happy to follow your recommendations. We will have double peer reviewed set of short papers.

Any other topics relevant for discussion.

We removed the line numbers, but if the intention is to publish this text, we should remove or update the names of keynotes and PC.
Thanks for supporting this workshop – we will explore the option to organise it as a regular event in the next few years.

It would be useful to distribute to all workshop handlers for social media on CLIB, which we can quote when publishing our updates.