Metadata labeling of INFN products

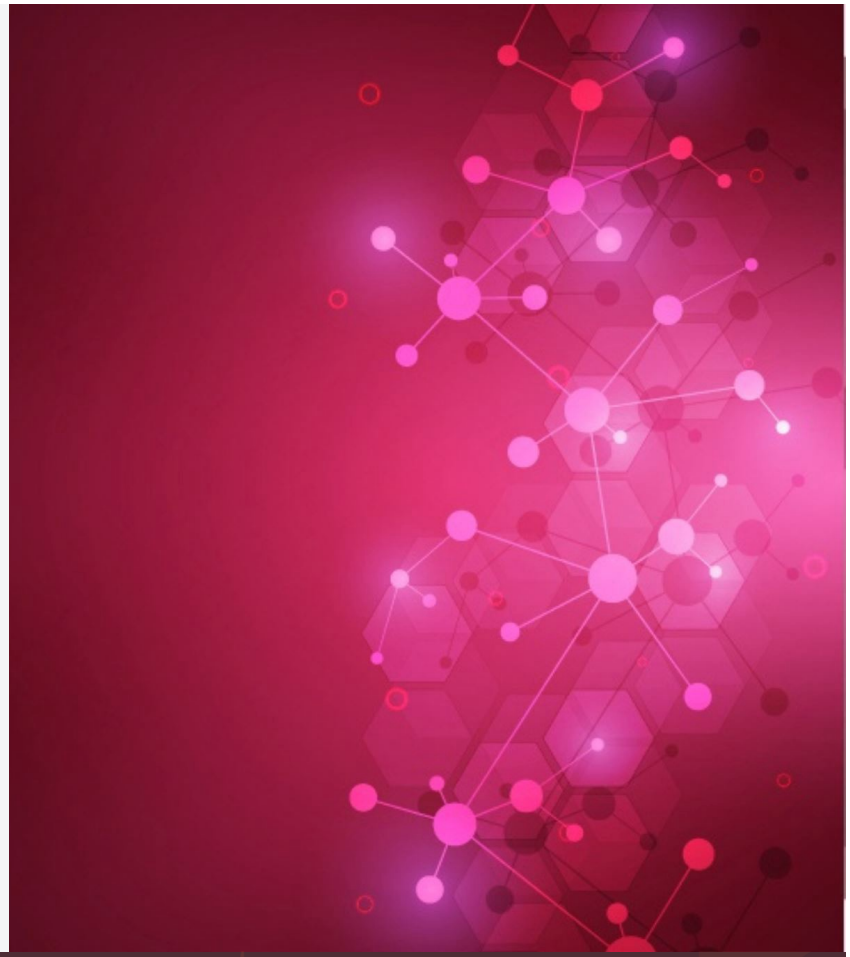
Speakers **A. Paoletti - M. Gattari**

Contributors **L. Sanelli - F. Serafini - M. D'Alessandro**

IT Direction - INFN Central Administration

in partnership with
*the INFN Evaluation Workgroup (**GLV**) and the INFN Open Science Workgroup (**GLOS**)*

DOI: **10.15161/oar.it/211866** - GenOA week – november 2024



Overview

- Introduction and context
- Issues
- Solutions
- Use case
- Remat

Introduction and context

- The Research Evaluation Process (VQR) is a quality enhancer of the research products organization lifecycle
  - Systematic process of products (and metadata) evaluation
  - Evaluation of author coverage according to VQR Rules
- From Research Quality Assessment to
  - Data quality assessment (content)
  - Process quality assessment (methods)
  - Tools quality assessment (infrastructure)

Issues: what undermines data quality

- Multi-step lifecycle and transitions across systems and actors
- Formal errors in parsing and validation across transitions
  - *Author - Publisher (ORCID not acquired or not mandatory)*
  - *Publisher – Aggregator (data model refactoring)*
  - *Aggregator – CRIS (data linking to internal databases)*
- Critical issues, if not mitigated, result in VQR penalties

Solutions

- Establish a virtuous cycle between users and algorithms
  - Data production and enrichment processes
  - Backoffice support activities
- Awareness and engagement of research staff
  - Researchers are both the **source** and the primary **users** of the data
- Development of semi-supervised control systems
  - Assisted data curation (AI/ML) for archive staff.

Use Case: Product Authors

- **Target**: identify the best allocation of research products while minimizing the penalties from the VQR call
- **Constraint**: catalog all relevant products in the CRIS and correctly link them to the research staff
- **Critical issue**: incorrect or missing links distort the data and the outcome of the evaluation !
- How can we maximize the accuracy of automated linking without compromising its effectiveness?

Use Case (continue)

- Active involvement of authors!
- Leverage ML mechanisms by 'learning' from user actions or threshold-based automated systems
- Utilize identifiers (e.g., ORCID, ROR) as much as possible instead of outdated terms (display name + affiliation)
- Question everything and perform asynchronous checks on the data
  - Highlight error situations by using threshold systems
  - Create control dashboards for back-office staff

Metadata consistency

We collect metadata from several sources.

Problem: metadata consistency, e.g.:

- **author's identity**:
  - Paolo Giovanni Rossi
- **aliases**:
  - Rossi, Paolo Giovanni ✓
  - Rossi, PG ✓
  - Grossi, P ✗
- **orcid**:
  - 0000-0001-2345-6789 ✓
  - 0000-0001-XXXX-YYYY ✗
- **affiliations**:
  - INFN Frascati Natl Labs, I-00044 Frascati, Roma ✓
  - INFN Sez, Lab Nazl Frascati, Rome ✓
  - Univ Siena, Dipartimento Fis, Pisa, Italy ✗

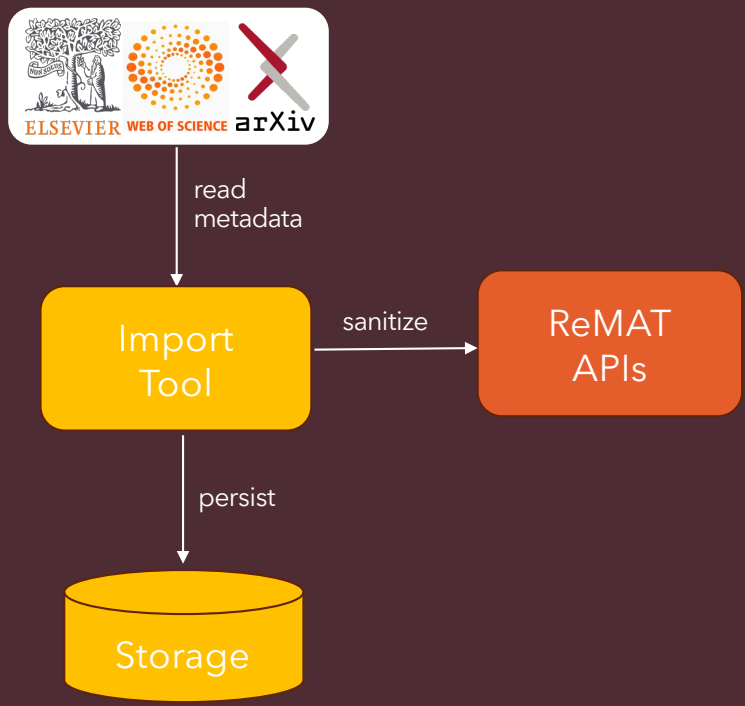
ReMAT

Research Metadata Analysis Tool

Open-source tool for **metadata validation** about authors and research products

Highlights:

- Automatic **reporting** of inconsistent data.
- Offline operations: **command line interface** for bulk analysis.
- Online operations: **web APIs** for data assessment.
- Written in python, customizable and extendible.



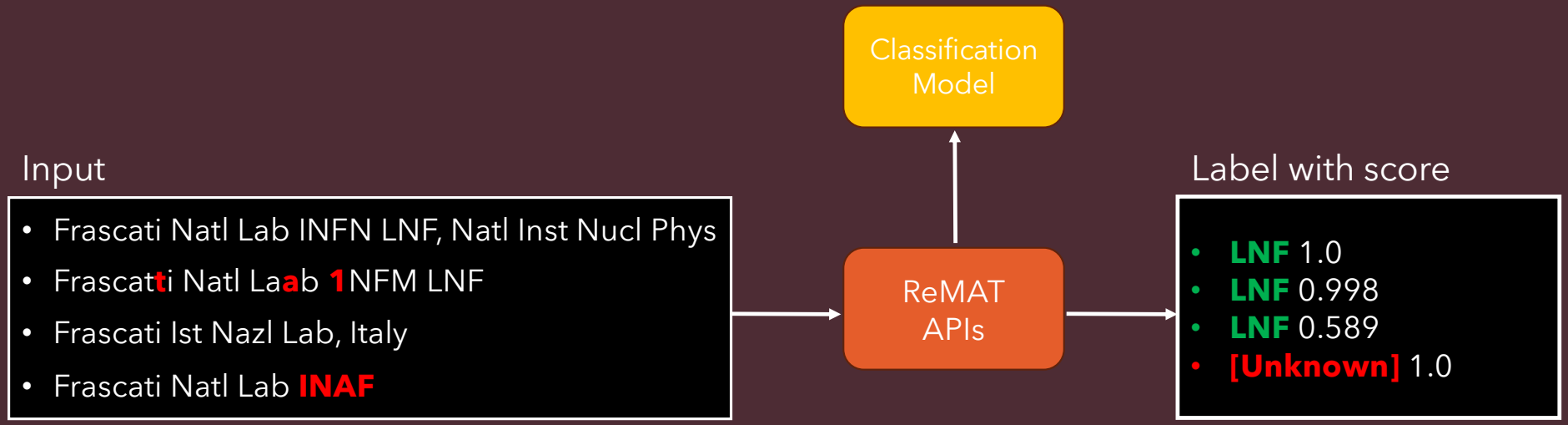
ReMAT

Author's affiliation classification

- Task: classify author's affiliation strings using a classification model.
- We have trained a (transformer-based) classification model over INFN author's affiliation strings.
- Training dataset:
  - ~6k positive samples:
    - **"INFN Frascati Natl Labs, I-00044 Frascati, Roma" -> LNF**
    - **"INFN Bari, Dept Phys, Bari, Italy" -> BA**
  - ~6k negative samples:
    - **"Univ Siena, Dipartimento Fis, Pisa, Italy" -> [Unknown]**
- Dataset augmented to ~400k samples by adding "smart" typos
- Training evaluation: 97% accuracy on test set

ReMAT

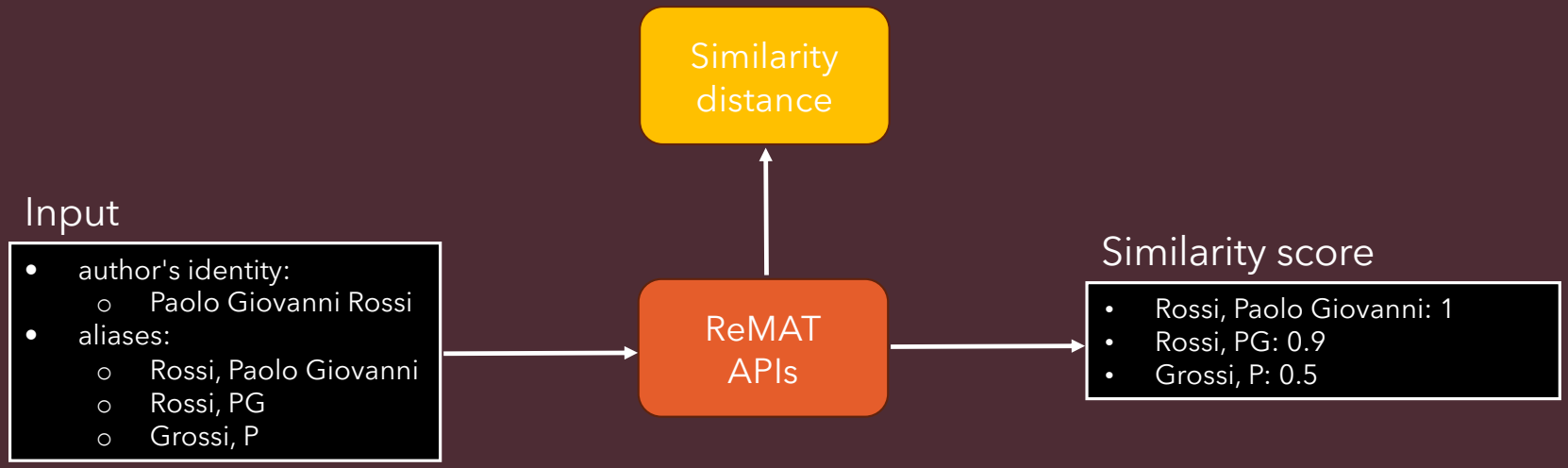
Author's affiliation classification



We are interested in recognizing INFN authors, but you can provide your own classifier, labels and threshold

ReMAT

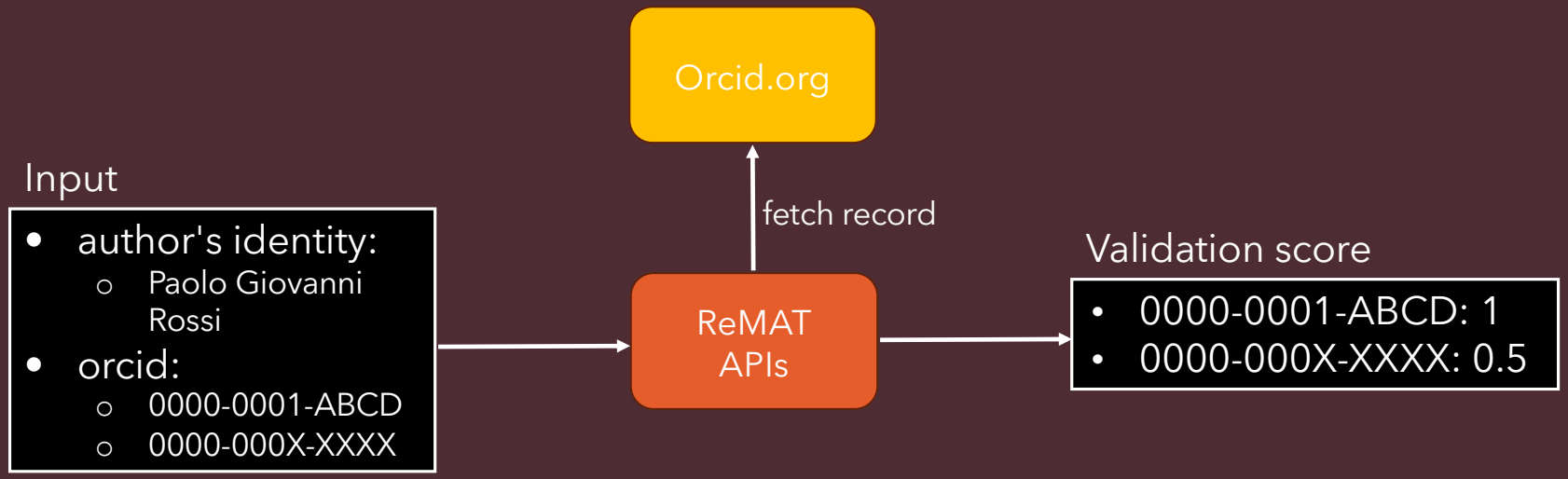
Author names similarity



We use the Jaro-Winkler distance to compute the similarity score, but you can provide your own metrics and thresholds

ReMAT

Orcid validation



We calculate similarity score between author's identity and Orcid.org name(s).
Here again you can configure your own metrics and thresholds.

Thank you !

antonello.paoletti
mauro.gattari
luca.sanelli
francesco.serafini
marzio.dalessandro
@infn.it