The Information Retrieval Experiment Platform (Extended Abstract)

IJCAI 2024

Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, **Niklas Deckers**, Simon Reich,
Janek Bevendorff, Benno Stein, Matthias Hagen, Martin Potthast

Friedrich-Schiller-Universität Jena, University of Glasgow, Leipzig University, ScaDS.AI,
Bauhaus-Universität Weimar, University of Kassel, hessian.AI

Invited extended abstract of our SIGIR 2023 Best Paper.

The Importance of IR in the Era of Generative Models

- ❑ With more and more AI systems (e.g. LLMs) becoming available, defining evaluation metrics and evaluating the systems becomes important and this is where IR is very strong

- ❑ Transferring methods from IR to AI:
  - Evaluation metrics
  - Modelling interactions between humans and systems
  - Generative models (LLMs, RAG, text-to-image models) behave like search engines, searching over an Infinite Index

Motivation: Common Problems of AI and IR

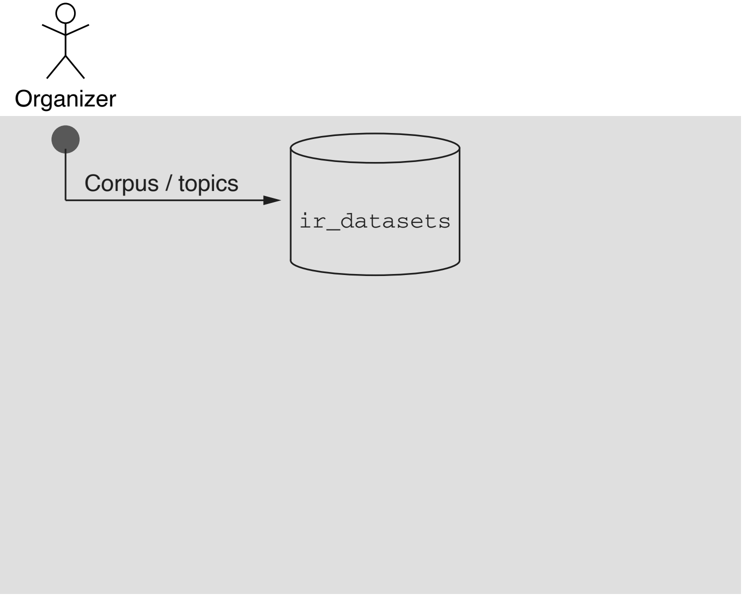
- Shared tasks and competitions are typically used to compare systems in domains like NLP, Computer Vision and IR

- Reported evaluation scores suffer from test data leakage

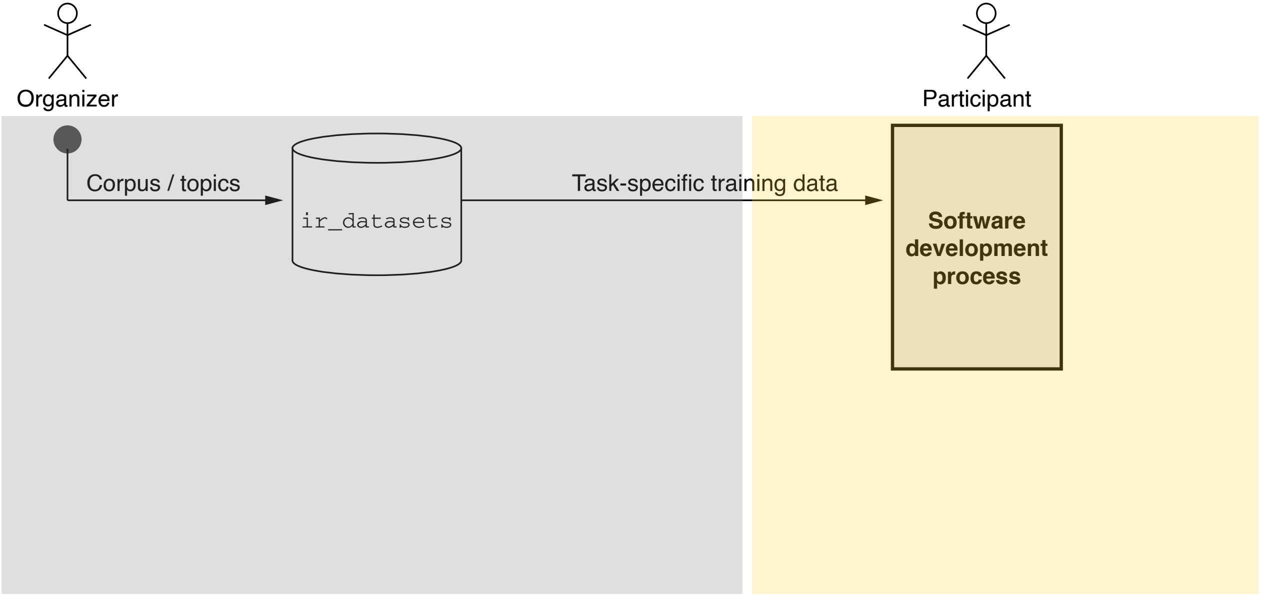
- AI has a reproducibility and replicability problem:
  - Blackbox models and API-only cloud models
  - Models with intransparent versioning and hidden updates
  - Local evaluation

- LLMs have become core component of IR systems

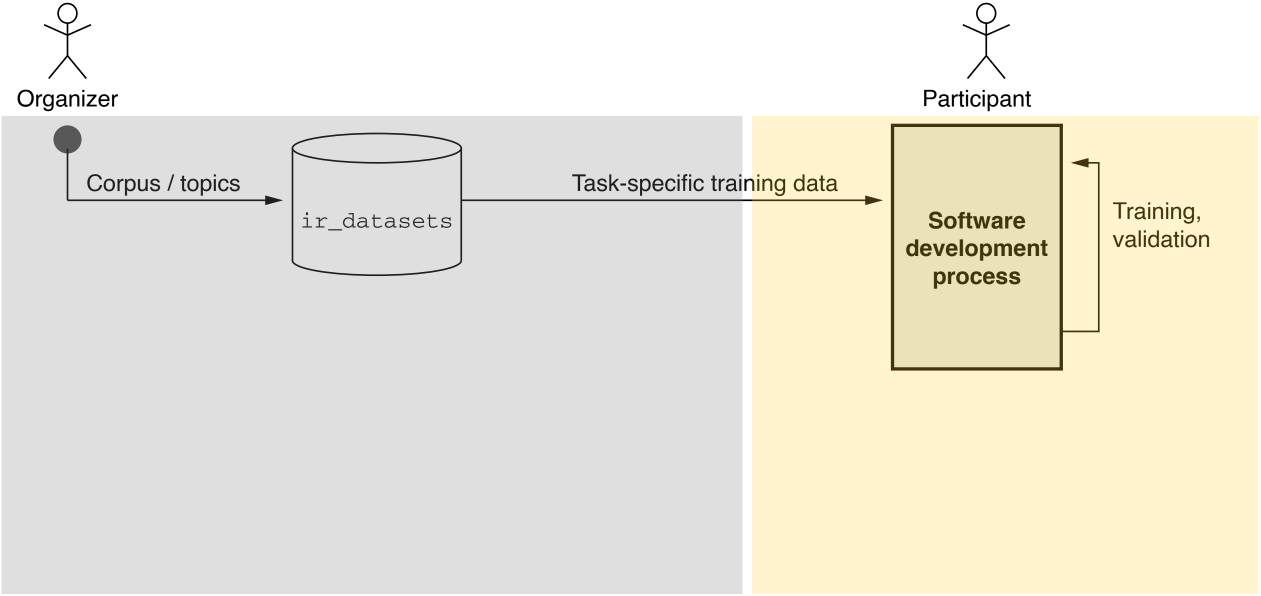
Approach: The Information Retrieval Experiment Platform



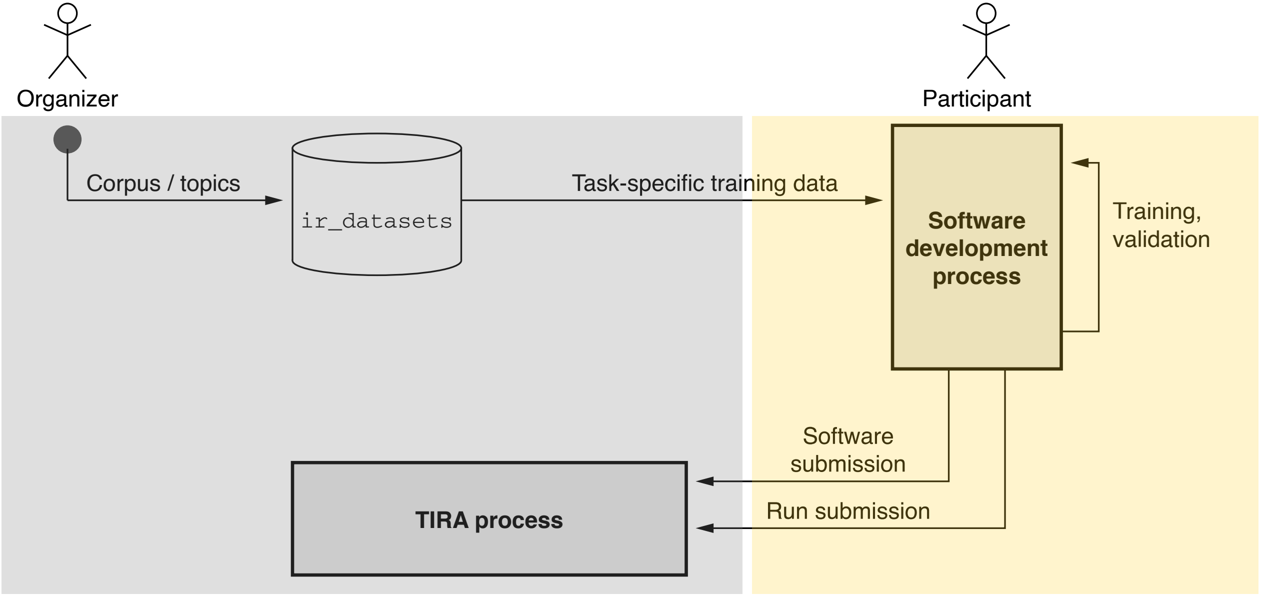
Approach: The Information Retrieval Experiment Platform



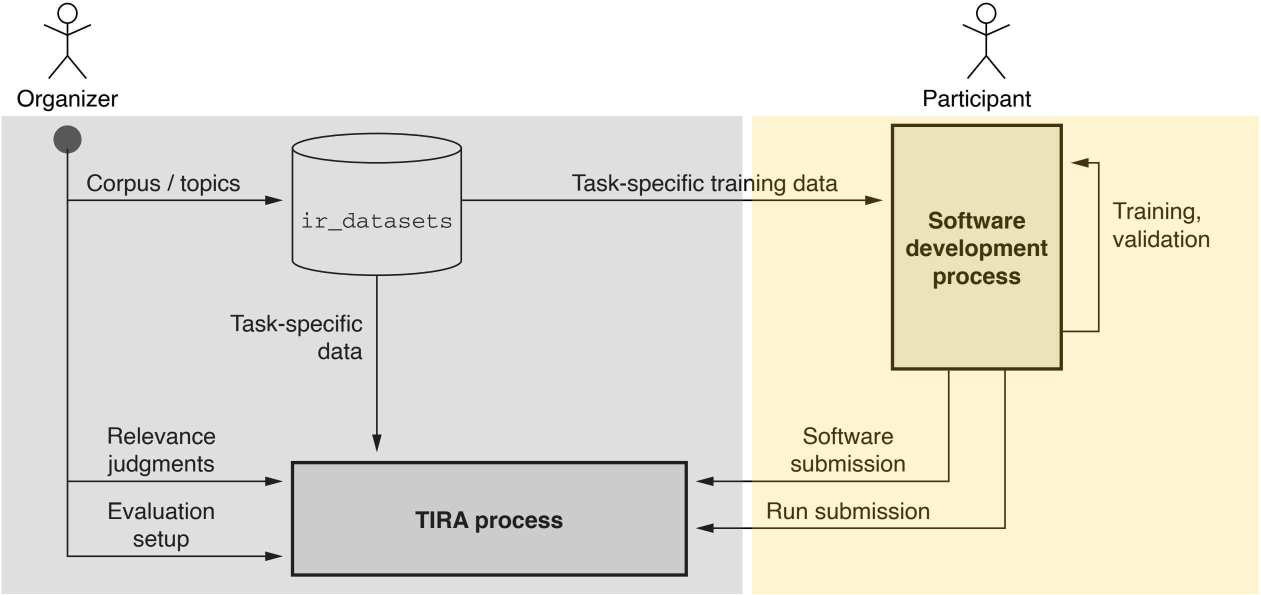
Approach: The Information Retrieval Experiment Platform



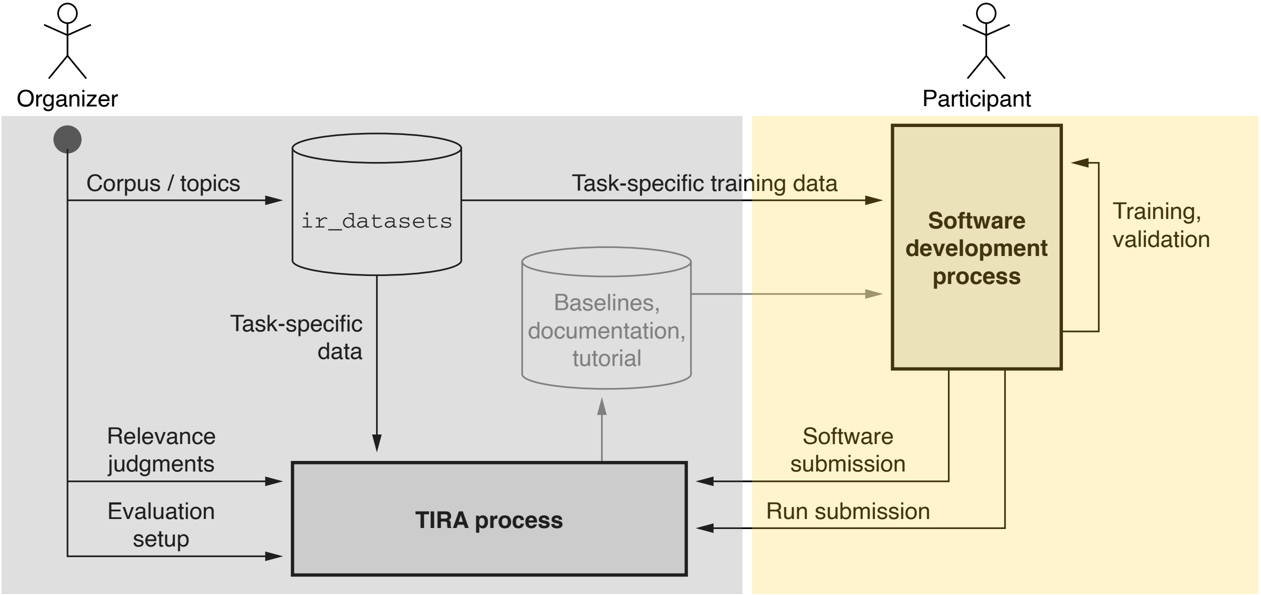
Approach: The Information Retrieval Experiment Platform



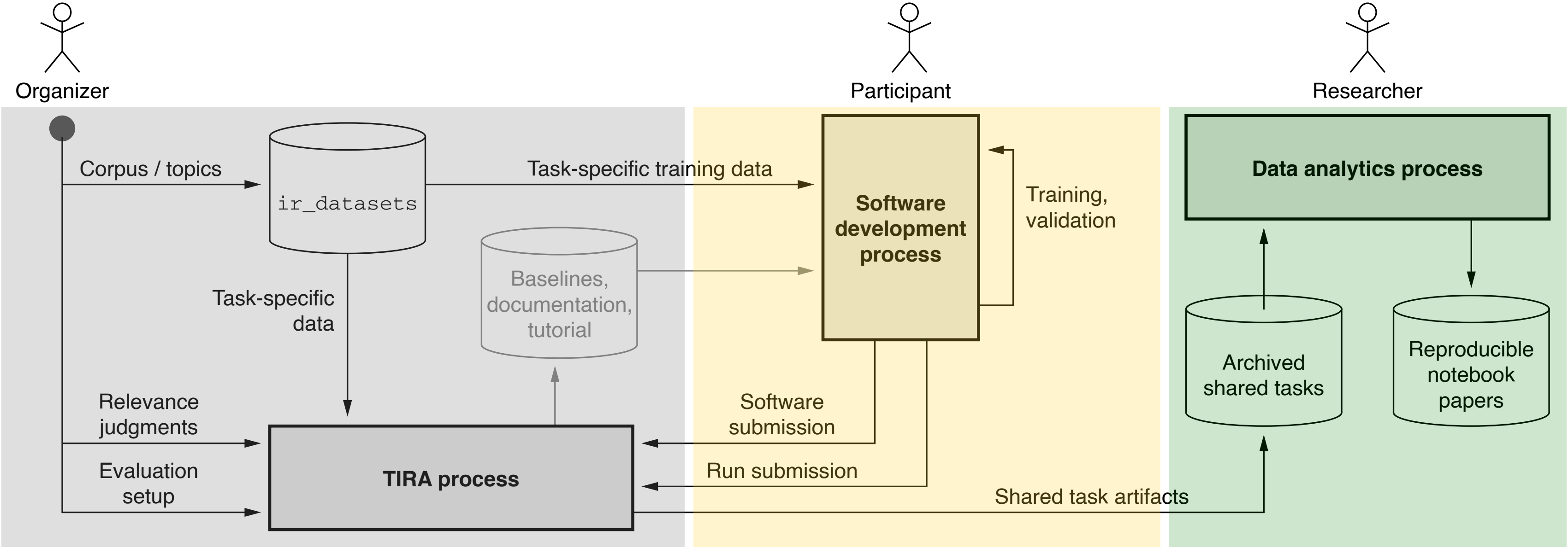
Approach: The Information Retrieval Experiment Platform



Approach: The Information Retrieval Experiment Platform



Approach: The Information Retrieval Experiment Platform



Advantages

- Sandbox architecture allows to keep test data secret,
  - preventing leakage and
  - allowing to execute software on datasets that are not (yet) to be shared publicly

- Participant software can be reused for further analysis/tasks/pipelines since everything is dockerized

- TIRA is compatible with evaluation scenarios beyond IR

- Supports. . .
  - Experiments with generative models
  - Loading models from Hugging Face Hub
  - GPU-based computations
  - LLM integration: Allows participants to use shared LLMs

Applications

❑ In IR: Integration of typical datasets and workflows from IR

❑ 50 baselines have been evaluated on 32 benchmarks

❑ Shared tasks in domains like NLP (e.g. PAN)

❑ Used in university courses

Applications

- In IR: Integration of typical datasets and workflows from IR
- 50 baselines have been evaluated on 32 benchmarks
- Shared tasks in domains like NLP (e.g. PAN)
- Used in university courses

Original SIGIR 2023 Best Paper