Technische Universität Berlin

Advanced Research Data Infrastructures

Project Advanced Research Data Infrastructures

Introduction Class
Summer Semester 2024

Outlines

- Who are we and you?
- Goals of Project Advanced Research Data Infrastructures (PJ ARDI)
- Presentation of Offered Projects
- Grading

Who are we?

- **Research Data Infrastructure Group**
  - Established 10/2023
  - http://tu.berlin/fdi

Who are we?

Prof. Dr. Sonja Schimmler

Previous

- University of the Federal Armed Forces Munich
  - PhD in Computer Science
- Technical University of Munich & Georgia Institute of Technology
  - Master in Computer Science

Current

- Fraunhofer FOKUS: Research Group Lead
- Technical University of Berlin: Guest Professor

Research Topics

- Digitalisation and Opening up of Science
- Research Data Infrastructures

Who are we?

Guest Advisors

- Dr. Muhammad Ahtisham Aslam, muhammad.ahtisham.aslam@fokus.fraunhofer.de
- Zongxiong Chen, zongxiong.chen@fokus.fraunhofer.de
- Damien Focard, damien.foucard@tu-berlin.de
- Sefika Efeoglu, sefika.efeoglu@tu-berlin.de
- Yue Zhang, yue.zhang@tu-berlin.de

Research Focus

Modern research on an internationally competitive level…

- …relies on more and more data
- …is no longer possible without digital support

Live Sciences

Natural Sciences

Engineering Sciences

Humanities and Social Sciences

This leads to a paradigm shift

Scientific progress will only be furthered, if…

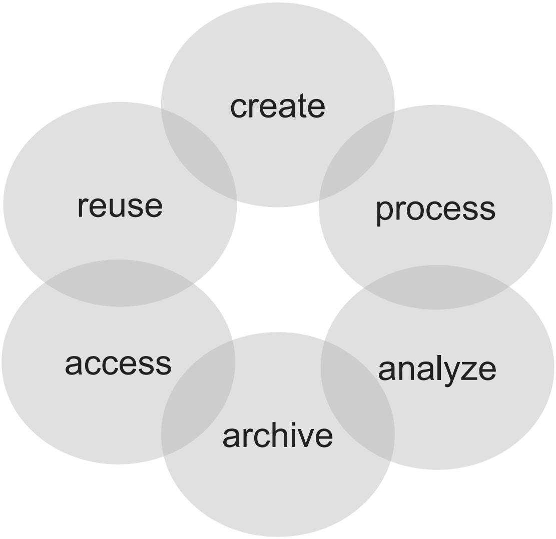
(1) …data becomes available at large scale

(2) …data is linked and machine-interpretable

(3) …specificities of the different disciplines are very carefully taken into account

Research Focus

- Meta portals to give unified access to data available in different sources
- Tools to analyze and process data and metadata



- The whole **research data life cycle** should be considered
- Publications, research data, and other **digital artifacts** (e.g. software) should be taken into account and should be properly represented
- Digital artifacts should be enriched with **metadata** and should be **linked** for contextualization
- The technical foundation is formed by **semantic web**, **linked data** and **knowledge graphs** on the one side, and **data science** and **artificial intelligence** on the other side.

Current Projects: Weizenbaum Institute

The **Weizenbaum Institute** is „the German Internet Institute“. It focuses on interdisciplinary digitalisation research.

Lead of the research group **Digitalisation and Science**

– What opportunities and risks arise from the digitalisation and opening up of science? What risks and opportunities arise from novel data infrastructures and tools?

– How can digital artifacts be best represented and interlinked?

– What data infrastructures and tools are well suited for interdisciplinary and data-intense research? How can these systems be improved?

Current Projects: BUA

The **Berlin University Alliance** is developing Berlin as a research location with international appeal.

Principal investigator of the project

Open Science by Design: New Methods for Research Data Infrastructures

- Analysis of current research processes
- Experimentation with new digital methods to support modern research

Principal investigator of the project

A Digital Research Space for the Berlin University Alliance

- Development of a meta portal to bundle different data sources
- Development of tools for searching, exploring and analysing research data

Current Projects: NFDI

The **German National Research Data Infrastructure** (NFDI) aims to systematically make available the valuable data sets from science and research.

Spokesperson of the consortium

NFDI for Data Science and Artificial Intelligence (NFDI4DataScience)

Co-spokesperson of the consortium

NFDI for Catalysis-related Sciences (NFDI4Cat)

Co-spokesperson of the initiative

Basic Services for the NFDI (Base4NFDI)

- Development of of basic services for the whole NFDI

Who are you?

Participants

- master students in one of the degree programs:
  - Computer Engineering, Computer Science, Elektrotechnik, Information System Management,
- will earn 9 ECTS credit points
  - the expected workload for one participant is 270 hours in the semester
  - the average per week is 18 hours

Why you take part in a project?

The topic can be the point from where

- do you (we) extend it to your Master Thesis topic or
- do you (we) extend it to a publication on ACM, IEEE. (Note: this requires more work after the project)

Learning outcome

Project Participants (m/f/d):

- will work in a team of around 3 or 4 students in a project
- specify, implement, test and document a software component
- analyze results (and deficiencies) of the project work within a written documentation
- practice presentation techniques while presenting three times during the project and expose the project plan, execution and results
- submit project results, including source code (e.g. Gitlab link), demo, documentation

Offered projects

Advisor: Zongxiong Chen (Fraunhofer FOKUS)

Topic: Boost Hyperparameter Tuning on domain specific LLM through Dataset

Offered projects

Advisor: Dr. Muhammad Ahtisham Aslam (Fraunhofer FOKUS)

Topic 1: Research Infrastructure and Research Publications Data Model (RIRPDM)

Topic 2: A Linked Open Data Based Approach to Track Mutual Citation Network in Research Publications

Research Infrastructure and Research Publications Data Model (RIRPDM): A Semantically Enriched Data Model for Representation and Reasoning of Research Infrastructure

Dr. M. Ahtisham Aslam | Summer term 2022

Why Semantically Enriched Research Infrastructure?

- *“Oil Based Economy”* □ *“Data Based Economy”*
- *Intelligent Analysis of Data* □ *Growth in Economy*
- *Research Data + Infrastructure* □ *Growth in Research Productivity*

  □ *Growth in Innovation*

  □ *Growth in All Sectors*

How to Develop Semantically Enriched Research Infrastructure?

- Research Infrastructure and Research Publications Data Model *(RIRPDM)*
- Research Infrastructure and Research Publications Data Ontology *(RIRPDO)*
- *RIRPDM + RIRPDO* □ *Information Extraction*
  □ *Semantic Web Technologies*
  □ *Linked Open Data & Knowledge Graphs*
  □ *Advanced Research Infrastructure*

A Linked Open Data Based Approach to Track Mutual Citation Network in Research Publications

Dr. M. Ahtisham Aslam | Project Open Distributed Systems | Summer term 2022

A Researcher’s Profile

- Number of citations
- Citation means how many times research work or a researcher is being mentioned by other researchers in their scientific work
- Citation Semantics
  - An article or researcher is cited positively or negatively
  - Is citing a scholarly article has real scientific link with the main document?
  - Is citation based on related work or some mutual understanding between scholars to artificially
  - improve the number of citations?

Semantic Technologies and Researcher’s Profile

- Knowledge model for Scholar’s Profile
- Data crawling and extraction from various Profile Networks
- Documenting Citation Matrix as RDF Datasets
- Querying Linked Data to Produce Bigger Knowledge Graphs

Offered projects

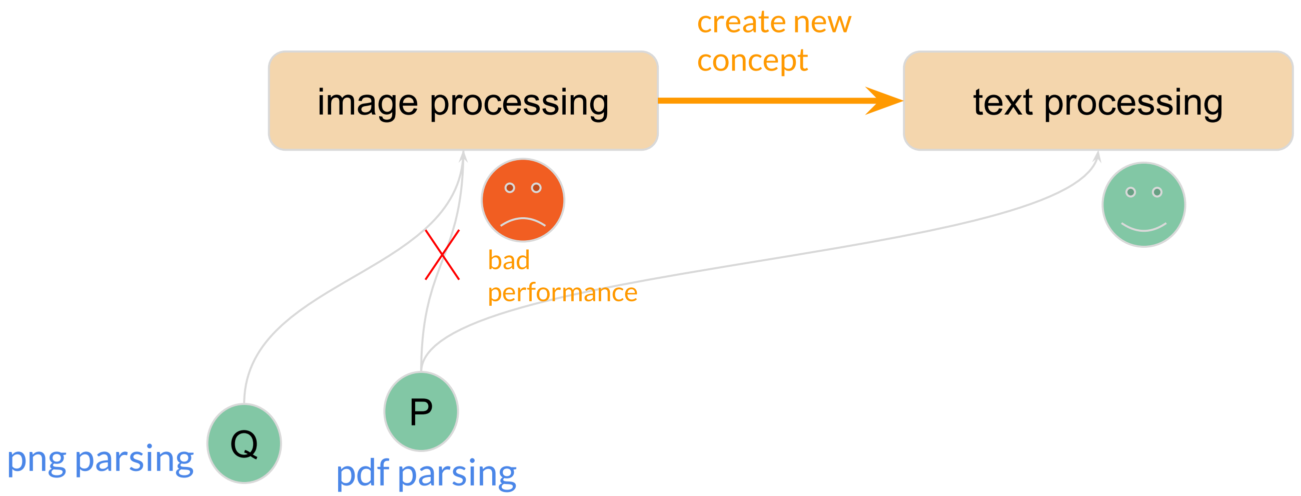
Advisor: Damien Foucard

Topic: Toward Autonomous Meta-Graphs

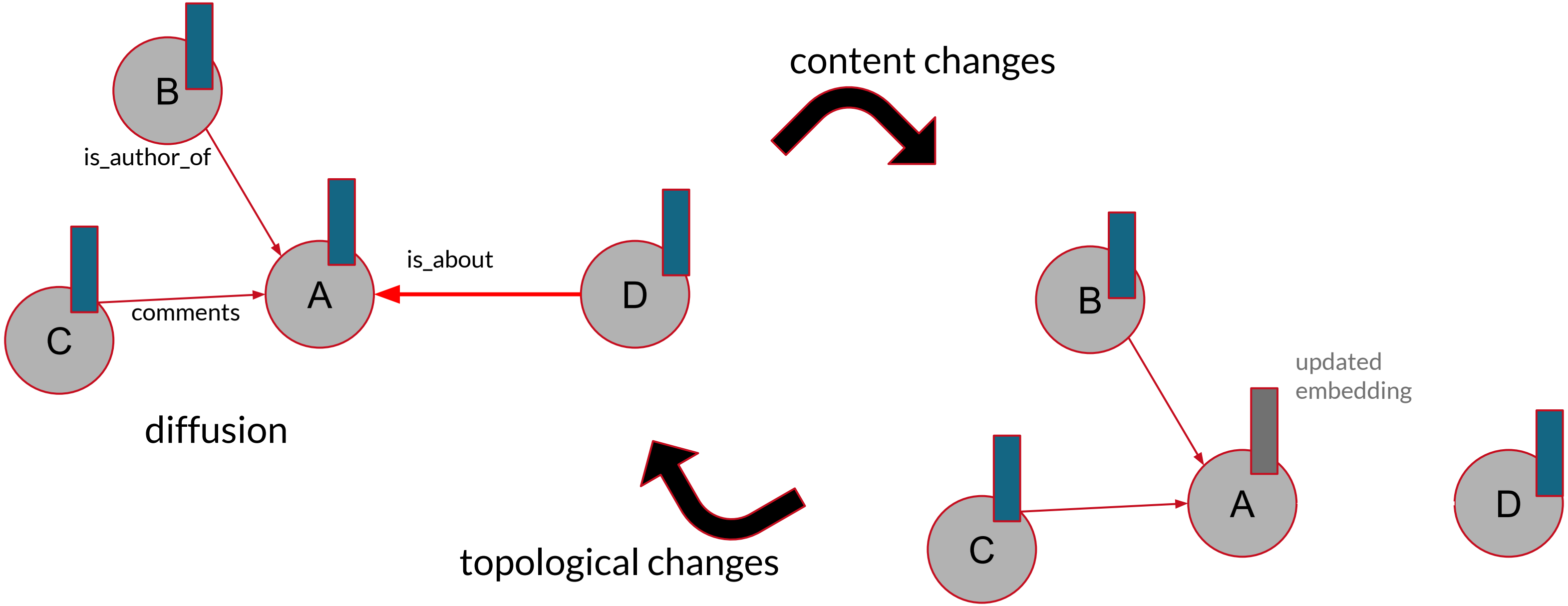
Toward Autonomous Meta-Graphs

PJ SS24 – Damien Foucard

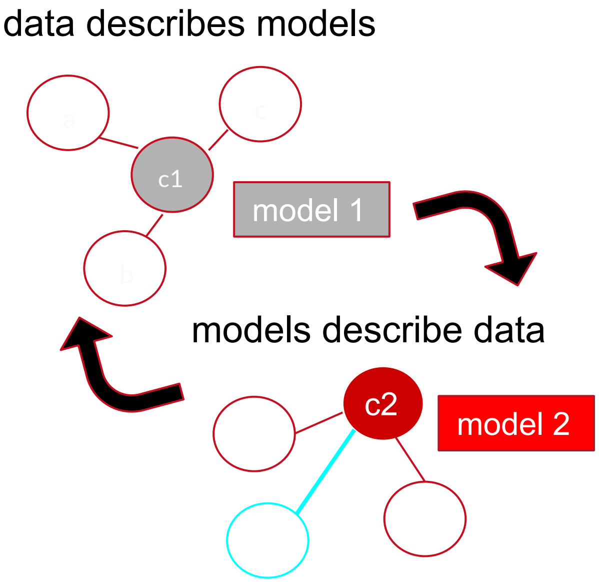
Dynamic Concept Creation



Passive Dynamic Graphs



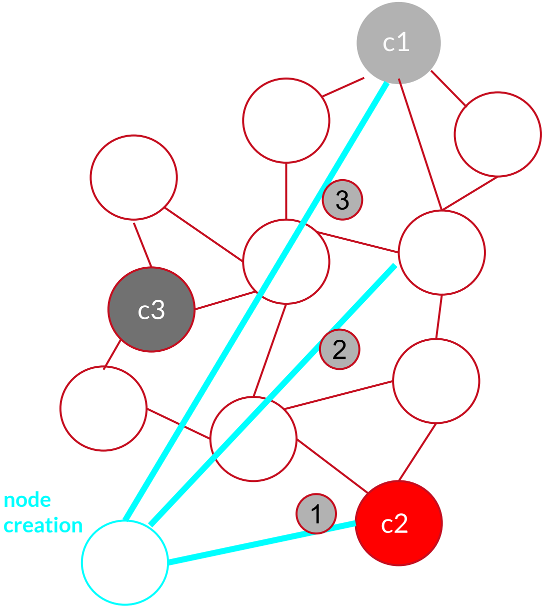
Model Mapping



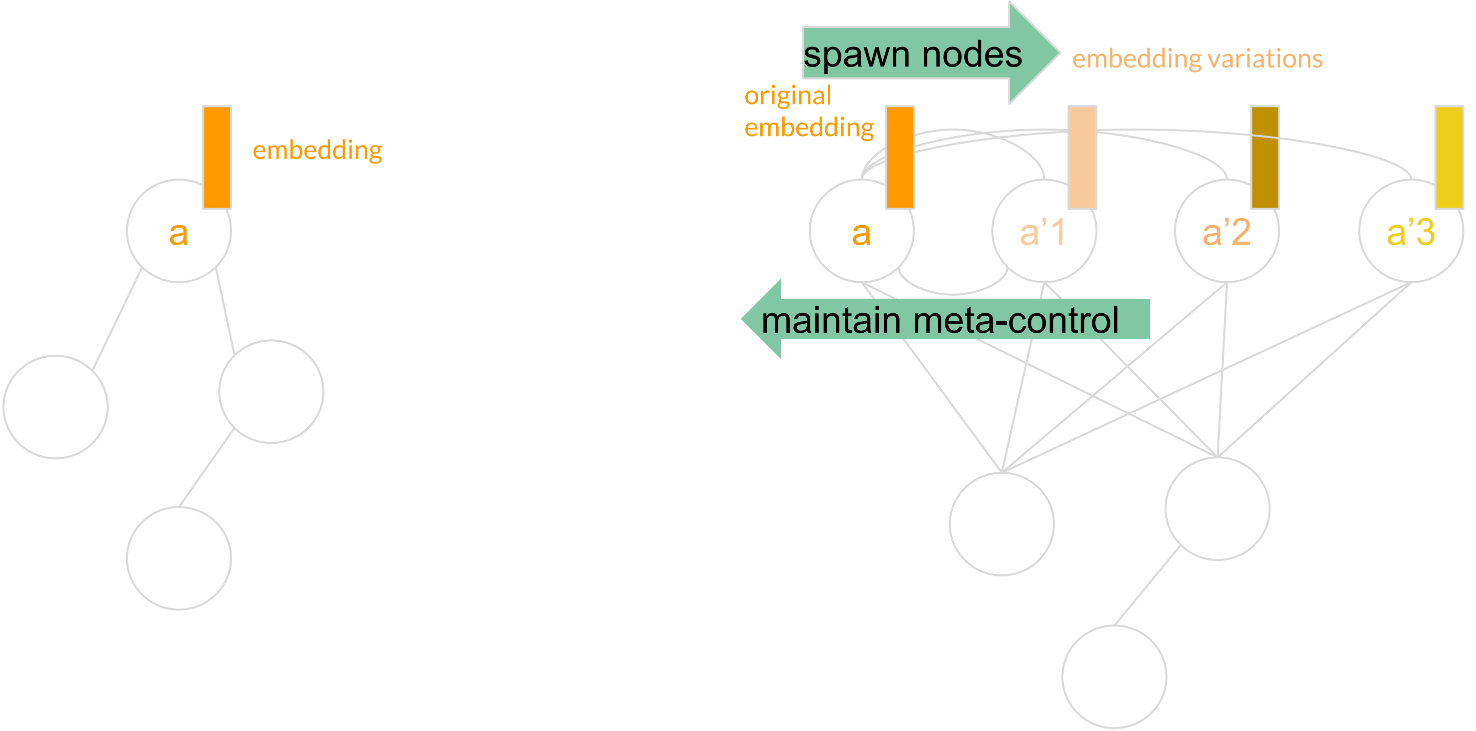
data describes models

models describe data

self-generating graph



How: Node Spawning



Offered projects

Advisor: Sefika Efeoglu

Topic: Text-to-SPARQL generation model utilizing LLMs for Scholarly Knowledge Graph Question Answering

Text-to-SPARQL generation model utilizing LLMs for Scholarly Knowledge Graph Question Answering

Sefika Efeoglu| Research Data Infrastructure | Summer term 2024

Introduction

The project aims to develop a scholarly Hybrid Knowledge Graph Question Answering (QA) that answers bibliographic natural language questions over multiple Knowledge Graphs and Text Sources e.g., DBLP KG, Wikipedia text, OpenaAlex KG etc. by leveraging multi-task fine-tuning of language models.

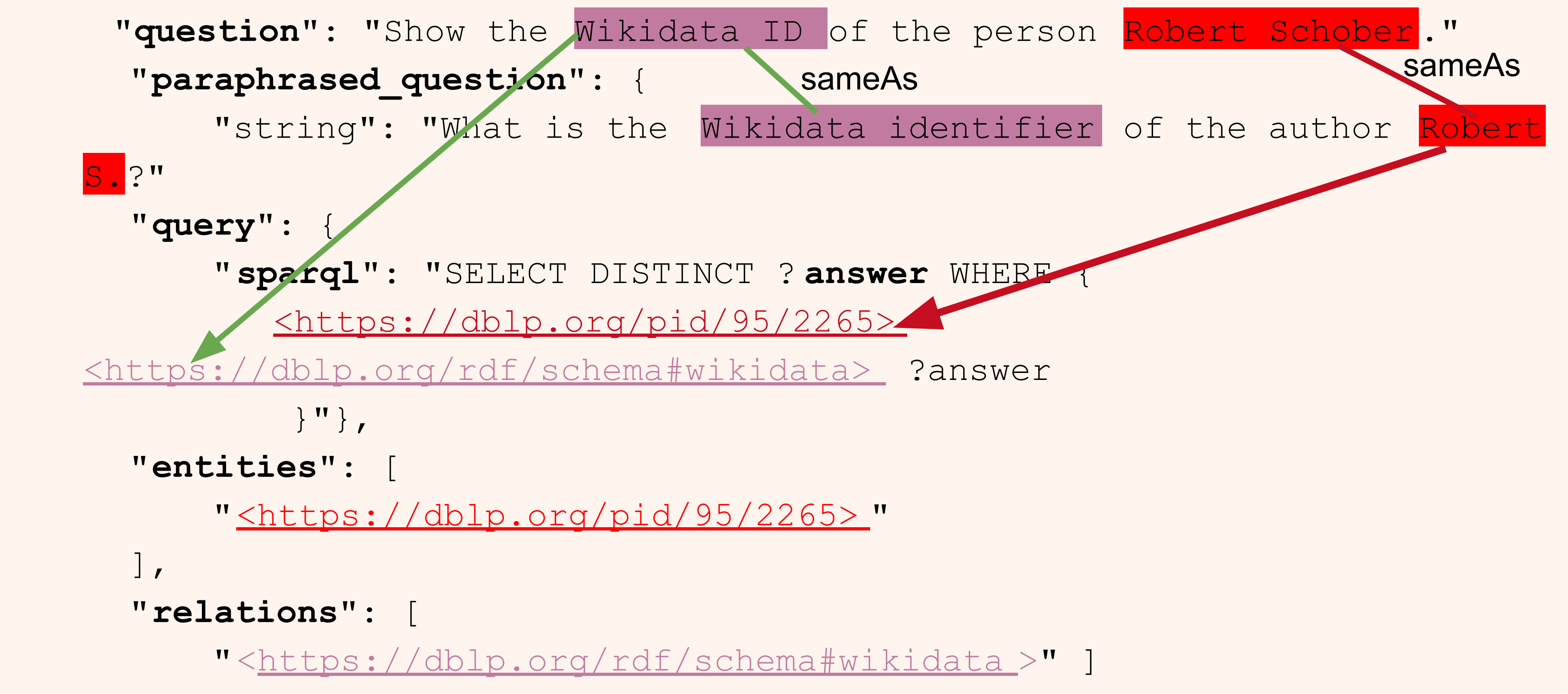
Tasks:

1. Entity Linking
2. SPARQL Generation

Dataset:

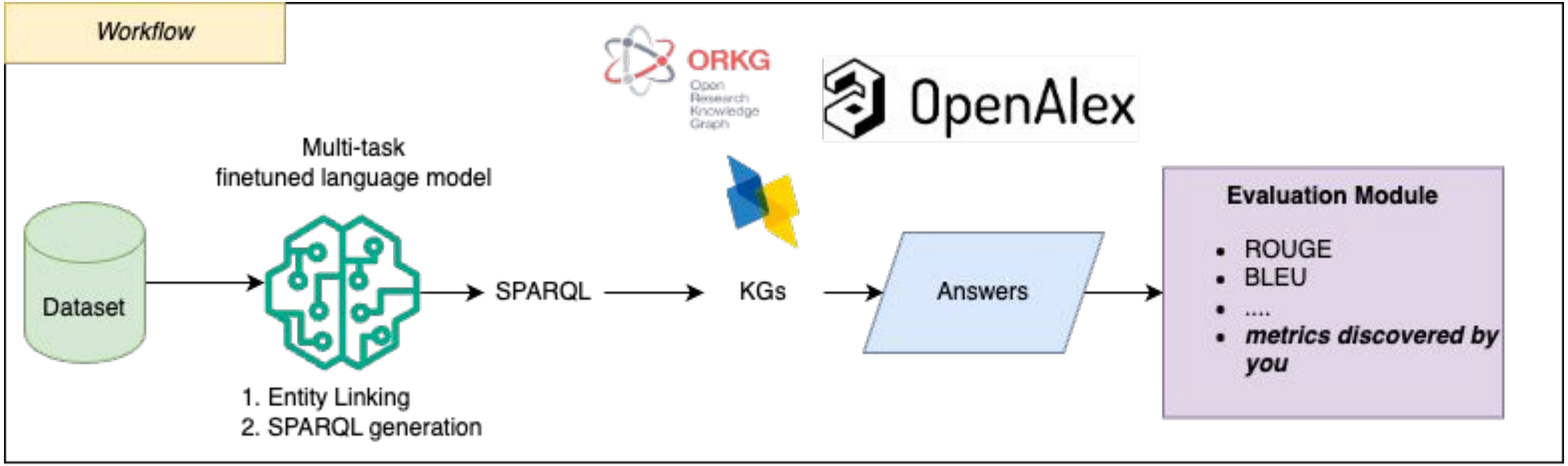
KGs like DBLP, Wikipedia Text, OpenAlex KG.

Dataset Overview



```
   "question": "Show the Wikidata ID of the person Robert Schober."
    "paraphrased_question": {
        "string": "What is the Wikidata identifier of the author Robert
S.?"
    "query": {
        "sparql": "SELECT DISTINCT ?answer WHERE {
           <https://dblp.org/pid/95/2265>
<https://dblp.org/rdf/schema#wikidata>  ?answer
             }"},
    "entities": [
        "<https://dblp.org/pid/95/2265>"
    ],
    "relations": [
        "<https://dblp.org/rdf/schema#wikidata>" ]
```

Methodology



Methodology

To fine-tune a language model for two tasks:

1. SFT, DPO Trainers
2. Parameter Efficient Fine Tuning (PEFT) : QLoRA, LoRA and other possible adapters will be added to the model.
3. Evaluation: implementation of corresponding text generation metric functions

Expected Outputs

1. Entities and their Unified Resource Identifiers (URIs) on KGs (output of entity linking part)
2. SPARQL generated from the natural language questions.
3. answers when SPARQL(s) is run on KGs

Offered projects

Advisor: Yue Zhang

Topic:Tracking the State-of-the-Art in Scholarly Publications

Tracking the State-of-the-Art in Scholarly Publications

Yue Zhang | Advanced Research Data Infrastructure | Summer 2024

Dataset

- Conventional ways looking for good papers:

  Leaderboards Searching, Reddit Topics, News or Blog, Research KG querying
- So many sources and so many papers
- In AI field, papers always has Tasks, Dataset, Metrics and corresponding Scores.
- SOTA? Dataset generated from ORKG [2] api functionings.
  - Available at: https://github.com/jd-coderepos/sota/
- Data Structure: Paper saved as tex format. Labels saved as json format.
  - unanswerable
  - *[{"LEADERBOARD": {"Task": "Change Point Detection", "Dataset": "TSSB", "Metric": "Relative Change Point Distance", "Score": "0.20066"}}]*
- Few-shot & Zero-shot Test-set
- Part of *SimpleText Track Paper: Improving Access to Scientific Texts for Everyone* [1]

Downstream

- Training on provided train set and evaluate on testset (few-shot)
- Handing the few-shot result to me and I will give you the score after April 28, 2024
- Join Codalab Competition (optional)

  CodaLab - Competition (upsaclay.fr)
- Zero-shot challenge (optional)
- Your own idea about how to play with this dataset.

Reference

- 1. Ermakova, Liana, et al. "CLEF 2024 SimpleText Track: Improving Access to Scientific Texts for Everyone." *European Conference on Information Retrieval*. Cham: Springer Nature Switzerland, 2024.

- Jaradeh, M. Y., Oelen, A., Prinz, M., Stocker, M., & Auer, S. (2019). Open Research Knowledge Graph: A System Walkthrough. In Lecture Notes in Computer Science (pp. 348–351). Springer International Publishing. https://doi.org/10.1007/978-3-030-30760-8_31

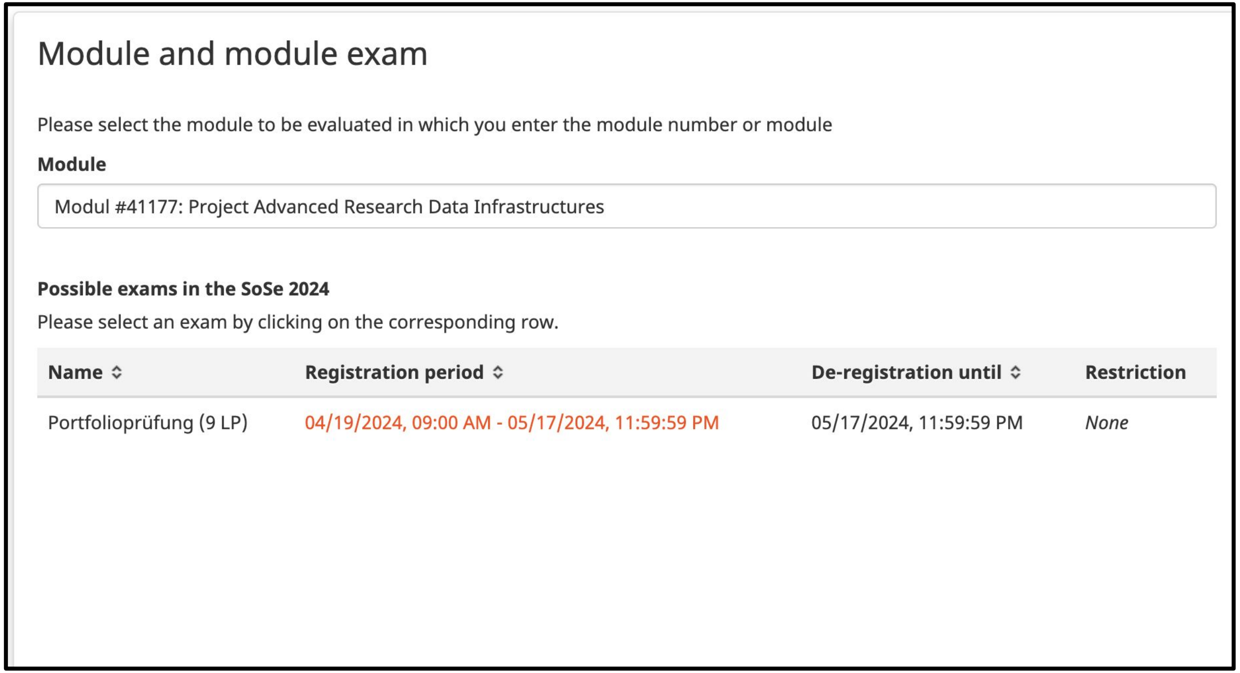
Tasks

Project Participants(m/f/d):

- **16.04** - Kick-off lecture. Introduction to Project ARDI
- **16.04** - begin of the project selection phase on ISIS (**at 5 pm**)
- **22.04** - end of the project selection phase (**at 10 am**)
- **23.04** - attend the first meeting with your supervisor
- **21.05** - give the first presentation and upload the presentation file to ISIS
- **25.06** - give the second presentation and upload the presentation file to ISIS
- **16.07** - give the final (3rd) presentation and upload the presentation file to ISIS
- **04.08** - hand-in documentation and project results (source code e.g. via Gitlab)

Grading

Exam registration - Prüfungsanmeldung



Module and module exam

Please select the module to be evaluated in which you enter the module number or module

Module

Modul #41177: Project Advanced Research Data Infrastructures

Possible exams in the SoSe 2024

Please select an exam by clicking on the corresponding row.

Name ⇕	Registration period ⇕	De-registration until ⇕	Restriction
Portfolioprüfung (9 LP)	04/19/2024, 09:00 AM - 05/17/2024, 11:59:59 PM	05/17/2024, 11:59:59 PM	*None*

Grading

Presentations

Presentations take at least 15 minutes and max (5 + members_count * 5) minutes.

Each group member should have an active part in the presentation.

- **First presentation**
  - Problem statement, paper review, responsibilities and schedule, next steps
- **Second presentation**
  - Wrap-up: problem statement, results and preliminary results, optional: first demo, update: schedule and next steps
- **Third (final) presentation**
  - Wrap-up: problem statement, final results and implementation details, final demo
- **All presentations**
  - keep in mind and present: project scope, design, work plan, team assignments
  - let the audience know about: project and risk management, technologies used, tool chains and deliverables

Grading

Documentation

- Project ARDI format (available ISIS course)
- 10-15 pages text (w/o table of content, images, reference list etc.)
- Scientific english
- Suggested structure:
  - Abstract
  - Introduction
  - Related work
  - Approach
  - Evaluation and discussion
  - Conclusion
  - References

Grading

Project Management

- the better and more flexible the team and project is managed
  → the better the probability for a very good solution
  - use an iterative and flexible method like SCRUM
  - the role of product owner is filled by the instructor
  - you might vote for a team leader
  - do a daily standup meeting (web conference)
  - plan weekly sprints and do sprint planning and sprint reviews (e.g. Gitlab)
  - use an issue tracker and a versioning tool (e.g. Gitlab)
- document your code
- review your code (pairwise) e.g. with merge requests
- be cooperative, flexible and self-reflective
- give and take feedback
- do not over-organize

Grading

Solution

- depends on project and advisor
- source code (e.g. on Gitlab)
- executable demo version (in case the prototype has several components, a docker image might be helpful)
- documentation (including a how-to, credentials)

Grading

How the grades are compound?

Total	Single	Description
30%	10%	1st presentation
	10%	2nd presentation
	10%	3rd presentation
30%	30%	documentation
40%	40%	project result, implementation

Grading

Points	100	>=95	>=90	>=85	>=80	>=75	>=70	>=65	>=60	>=55	>=50	<50
Grade	Max	1,0	1,3	1,7	2,0	2,3	2,7	3,0	3,3	3,7	4,0	5,0
ECTS	A	A	B	B	C	C	C	D	E	E	E	F

Further Reading

- TU Berlin, ISIS

  https://isis.tu-berlin.de/course/view.php?id=38049
- TU Berlin, Gitlab

  https://git.tu-berlin.de/
- TU Berlin, Qispos

  https://www.pruefungen.tu-berlin.de/menue/qispos/