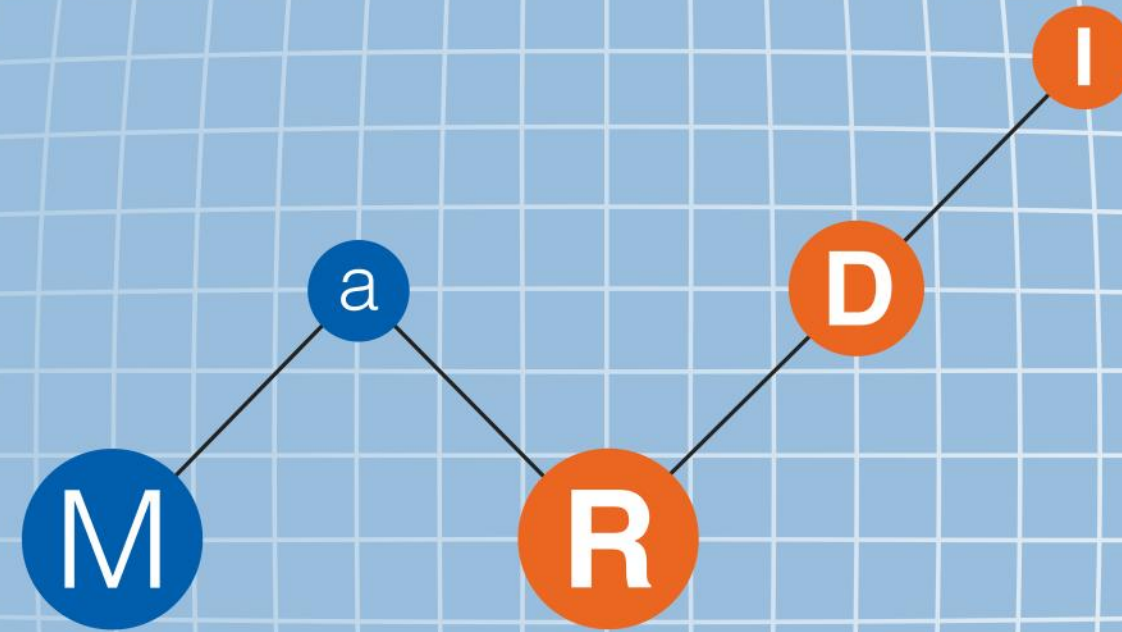


MaRDI

Mathematical Research Data Initiative

LLM-Assisted Extraction of Mathematical Model Metadata for FAIR Reuse



Jochen Fiedler¹, Felix Niclas Kreutz¹, Christine Biedinger¹, Michael Burger¹,
Dominik Göddeke², Thomas Koprucki³, Marco Reidelbach⁴, Aurela Shehu³,
Björn Schembera², Burkhard Schmidt³, Anita Schöbel¹, Jörg Schlötterer⁵, Marcus Weber⁴

¹Fraunhofer ITWM, ²Universität Stuttgart, ³WIAS Berlin, ⁴ZIB, ⁵Universität Marburg

Motivation: Create central repository for mathematical models

Challenges

- R&D teams often rely on mathematical models and concepts
- Models are often scattered across manifold publications and internal documents
- Finding the right model for a new research project (including assumptions, scope, required quantities, and references) is time-consuming
- Leads to fragmented knowledge and missed opportunities

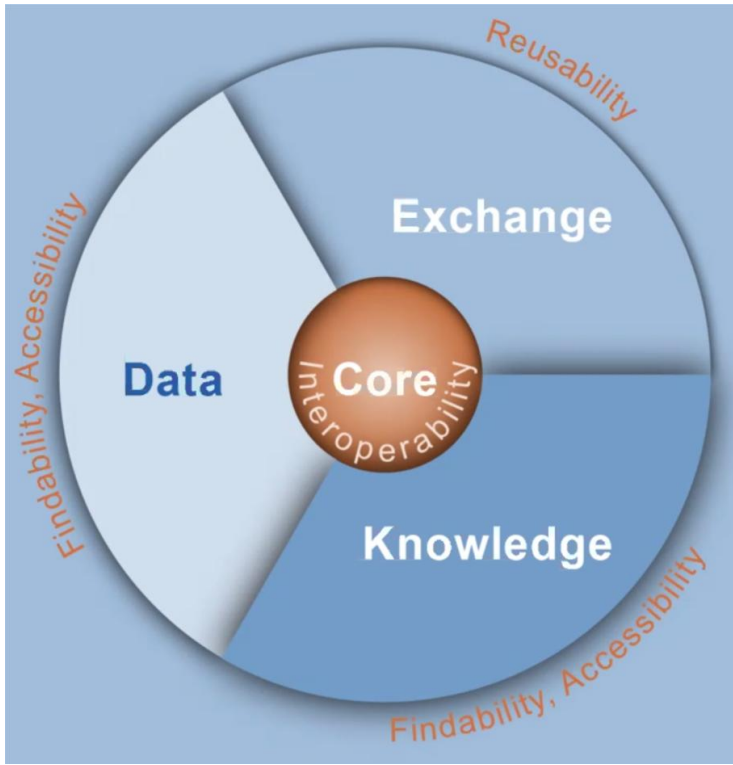
Goals

- Improve discoverability of models (including versions, variants, and application domains)
- Provide transparency and thorough documentation of necessary formulations and assumptions
- Connect and link research questions and domains to specific models
- Enable easy search, maximum synergies and unique identification

→ **Making mathematical models FAIR!**

MaRDI aims

- Develop robust Mathematical Research Data Infrastructure
- Set standards and confirmable workflows for certified Mathematical Research Data
- Provide services to both the mathematical and wider scientific community



Services

MediaWiki Math Search Extension
Tool/Application

The MaRDI portal team adjusted the extension for semantic formula search in the knowledge graph. The extension is used b...

[Discover more](#)

MaRDI Packaging System
Tool/Application

MaPS helps researchers create and publish software runtimes, as well as deploy and run software inside published runtime...

[Discover more](#)

MaRDI Knowledge Graph Database

The MaRDI Knowledge Graph connects over 5 million mathematical items by more than 500 million relationships from various...

[Discover more](#)

MaRDMO
Tool/Application

MaRDMO is a plugin designed to streamline the documentation of workflows. Primarily utilized for Model-Simulation-Optimi...

[Discover more](#)

MaRDI Help Desk
Outreach | Support/Consulting

The MaRDI Help Desk is your first entry point to MaRDI services, support, and training. Mathematical data consultant Chr...

[Discover more](#)

MaRDI Open Interfaces
Tool/Application

Software that connects different numerical packages together. Users can invoke numerical solvers written in one programm...

[Discover more](#)

mlr3
Tool/Application

mlr3 is an open-source machine learning framework in R that provides a unified interface for training, evaluating, and b...

[Discover more](#)

MathAlgoDB Knowledge Graph for Scientific Computing
Tool/Application

Algorithms are the main building blocks of scientific computing. MathAlgoDB is a knowledge graph with an underlying onto...

[Discover more](#)

MaRDIFlow
Tool/Application

This computational framework abstracts multi-layered components from FAIR computational experiments through an input/out...

[Discover more](#)

MathModDB
Database

MathModDB is a database of mathematical models developed by the Mathematical Research Data Initiative (MaRDI). MathModDB...

[Discover more](#)

mrDI File Format
Tool/Application

The mrDI file format is a JSON based file format with the necessary structure for saving and loading common types among ...

[Discover more](#)

Community - Graphical Modelling and Causal Inference
Curated Collection

On this platform, we curate and present topical datasets, dataset collections, and metadata...

[Discover more](#)

MaRDI Station
Outreach | Tool/Application

The MaRDI station offers an educational, gamified approach to research data management. It comes in two versions: a port...

[Discover more](#)

Best Practices
Support/Consulting

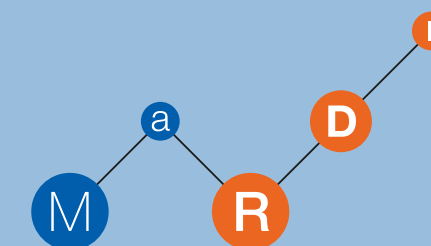
MaRDI offers support and consultancy for making your own mathematics FAIR. One example is the project "small phylogenet...

[Discover more](#)

MaRDI Knowledge Graph Query Service
Web application | Tool/Application

The MaRDI Knowledge Graph Query Service is a webservice that allows to query the MaRDI Knowledge Graph using SPARQL. For...

[Discover more](#)



The structure of MathModDB

Free Fall models as mathematical research data

- Simple model without air drag: $\dot{v} = g$
- More complex model with air drag: $\dot{v} = g - \frac{\rho C_D A v^2}{2m}$

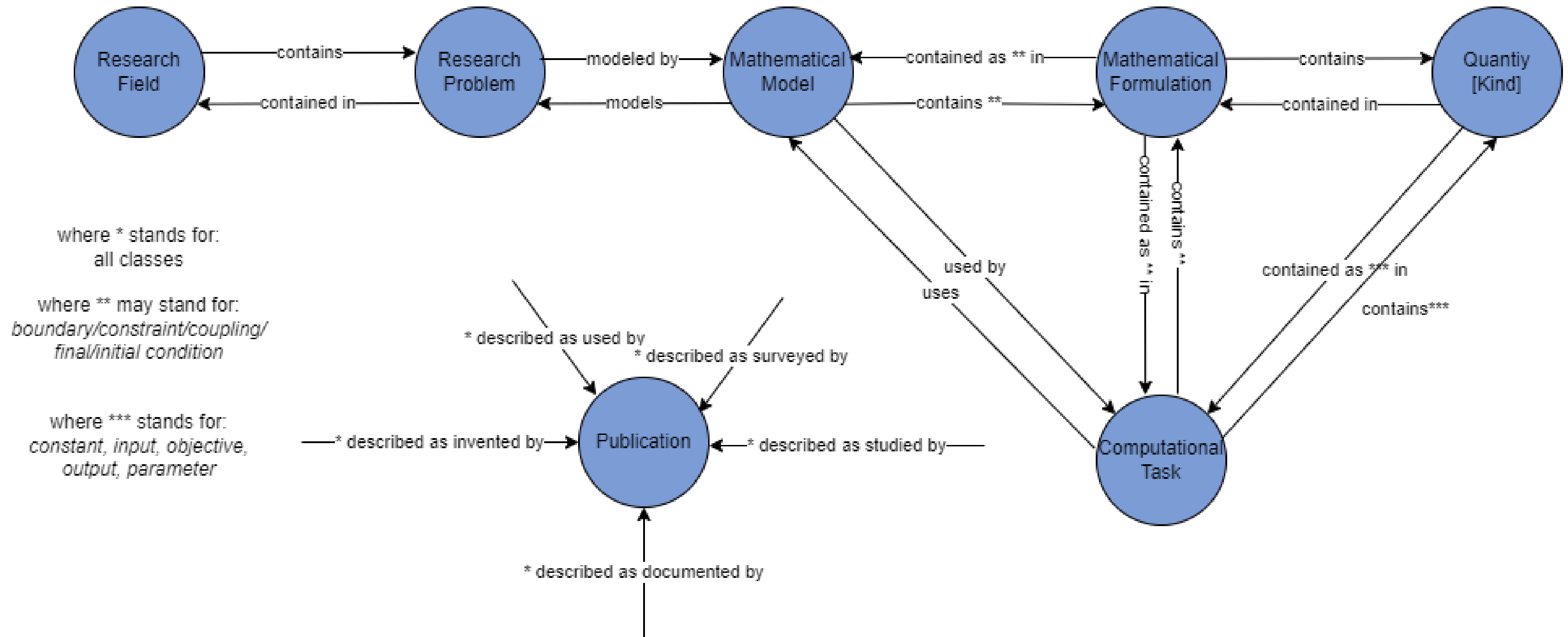
What and how to store such a model in a meaningful way?

→ Utilization of ontologies!



Sir Isaac Newton (1666)

Version 1.0.0 published in February



The free fall model in our ontology

Mathematical Model: Free fall with air drag

Mathematical Formulation: $\dot{v} = g - \frac{\rho C_D A v^2}{2m}$

Quantities: Free Fall Velocity v
Gravitational acceleration g
Density of air ρ
Drag coefficient C_D
Cross section A of the apple
Mass of the apple m

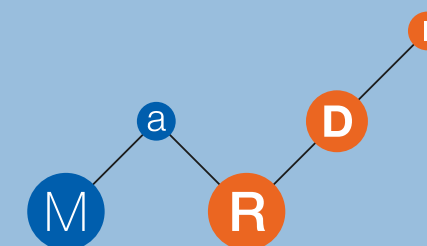
Computational Task: Calculate free fall time

Research Field: “Pomology”

Research Problem: Gravitational effects on fruit

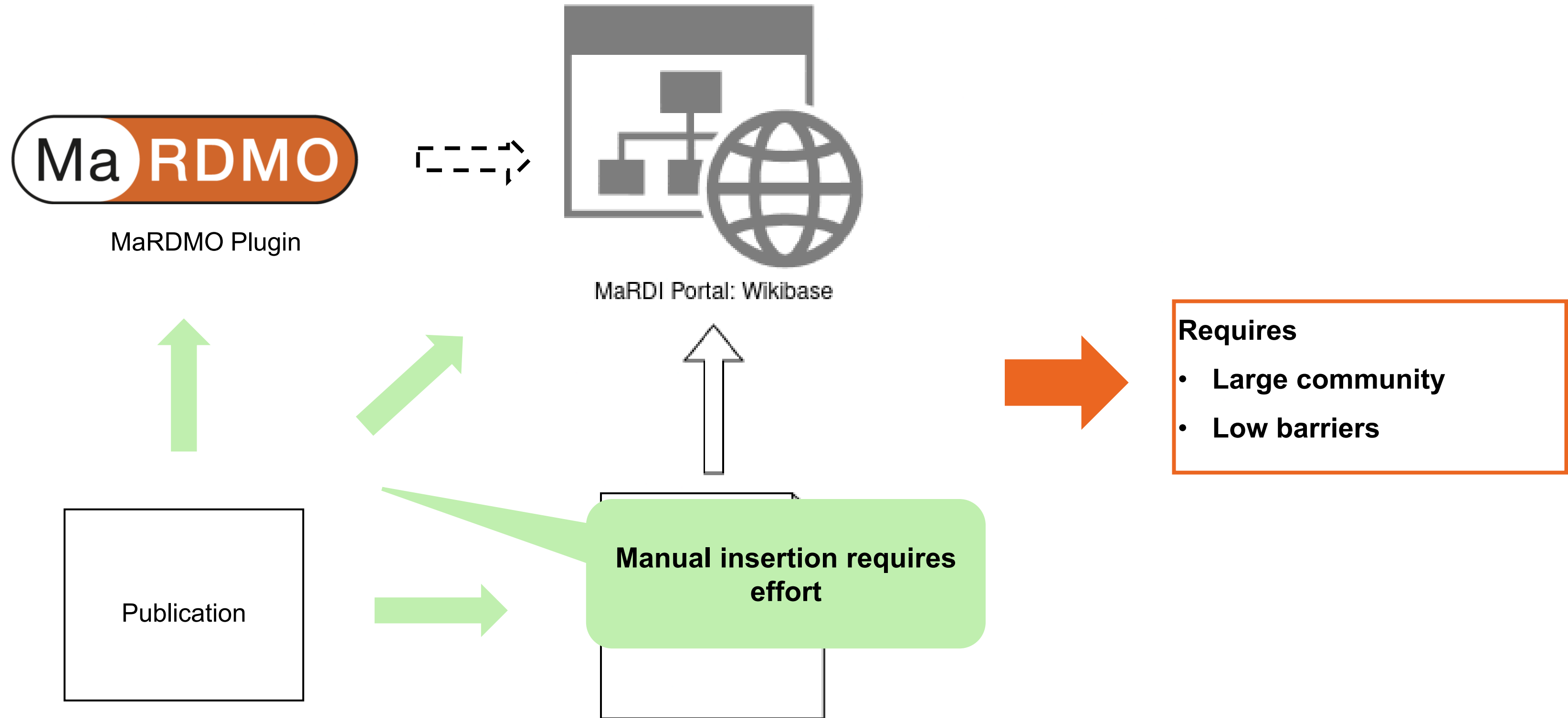


Sir Isaac Newton (1666)



Utilizing LLMs to fill knowledge graph

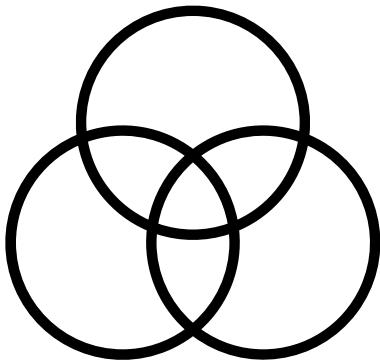
So far: Data needs to be inserted manually



Idea: Utilize LLMs to populate database



Publication
in Pdf



Triplets

Line pool generation11

pool is, the more complex it is to solve the **line planning model** in Phase 2. Hence, we require the number of **lines** in the **line pool** to be bounded by a value K to ensure tractability of Phase 2. To ensure that the **line pool** is ‘good enough’ is much harder. A minimal requirement for a good **line pool** is that it should at least be *feasible* for the **line planning problem** in Phase 2. Since there are so many different models for **line planning** it is unrealistic to guarantee feasibility for all possible models. However, the **edge frequency constraints** (1) are a minimal requirement which is used in most line planning models, and which we hence consider for designing a **line pool**.

We can now define the **line pool generation problem** for given K as follows.

(LPool) Given the PTN $= (V, E)$, lower and upper frequency bounds $f_e^{\min} \leq f_e^{\max}$ for all $e \in E$, and the set \mathcal{L}^c of all cycle free paths in the PTN, find a set $\mathcal{L} \subseteq \mathcal{L}^c$ such that $|\mathcal{L}| \leq K$ and such that there exist $f_l \in \mathbb{N}_0$ for all $l \in \mathcal{L}$ with

$$f_e^{\min} \leq \sum_{l \in \mathcal{L}: e \in l} f_l \leq f_e^{\max} \quad \forall e \in E.$$

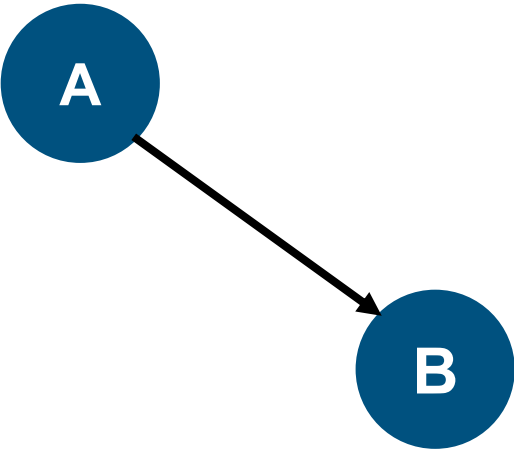
Using binary variables

$$\bar{x}_l = \begin{cases} 1 & \text{if line } l \text{ is chosen to be in the line pool} \\ 0 & \text{otherwise} \end{cases}$$

and variables $f_l \in \mathbb{N}_0$ to ensure the existence of a line concept, **LPool** can be formulated as the following integer program. To this end, let $M \geq \max_{e \in E} f_e^{\max}$.

$$\text{Find } \bar{x}_l \tag{2}$$
$$\text{s.t. } \sum_{l \in \mathcal{L}^c} \bar{x}_l \leq K \tag{3}$$
$$\sum_{l \in \mathcal{L}^c} f_l \geq f_e^{\min} \quad \forall e \in E \tag{4}$$
$$\sum_{l \in \mathcal{L}^c} f_l \leq f_e^{\max} \quad \forall e \in E \tag{5}$$

„Things not
Strings“



(Semi)-automatic data extraction lowers
barriers

Preliminary:

Extract textual information visually from Pdf utilizing Nougat (Neural Optical Understanding for Academic Documents)

Design: RoBERTa body with four different heads

Multitask-Learning Workflow:

- Knowledge Graph Embedding

Preliminary:

Extract textual information visually from Pdf utilizing Nougat (Neural Optical Understanding for Academic Documents)

Design: RoBERTa body with four different heads

Multitask-Learning Workflow:

- Knowledge Graph Embedding
- Detect entities from available classes (research problem, mathematical model, ...)



pool is, the more complex it is to solve the line planning model in Phase 2. Hence, we require the number of lines in the line pool to be bounded by a value K to ensure tractability of Phase 2. To ensure that the line pool is 'good enough' is much harder. A minimal requirement for a good line pool is that it should at least be *feasible* for the line planning problem in Phase 2. Since there are so many different models for line planning it is unrealistic to guarantee feasibility for all possible models. However, the edge frequency constraints (1) are a minimal requirement which is used in most line planning models, and which we hence consider for designing a line pool.

We can now define the line pool generation problem for given K as follows.

(LPool) Given the PTN= (V, E) , lower and upper frequency bounds $f_e^{\min} \leq f_e^{\max}$ for all $e \in E$, and the set \mathcal{L}^0 of all cycle free paths in the PTN, find a set $\mathcal{L} \subseteq \mathcal{L}^0$ such that $|\mathcal{L}| \leq K$ and such that there exist $f_l \in \mathbb{N}_0$ for all $l \in \mathcal{L}$ with

$$f_e^{\min} \leq \sum_{\substack{l \in \mathcal{L}: \\ e \in l}} f_l \leq f_e^{\max} \quad \forall e \in E.$$

Using binary variables

$$x_l = \begin{cases} 1 & \text{if line } l \text{ is chosen to be in the line pool} \\ 0 & \text{otherwise} \end{cases}$$

and variables $f_l \in \mathbb{N}_0$ to ensure the existence of a line concept, LPool can be formulated as the following integer program. To this end, let $M \geq \max_{e \in E} f_e^{\max}$.

$$\text{Find } x_l \tag{2}$$

$$\text{s.t. } \sum_{l \in \mathcal{L}^0} x_l \leq K \tag{3}$$

$$\sum_{\substack{l \in \mathcal{L}^0: \\ e \in l}} f_l \geq f_e^{\min} \quad \forall e \in E \tag{4}$$

$$\sum_{\substack{l \in \mathcal{L}^0: \\ e \in l}} f_l \leq f_e^{\max} \quad \forall e \in E \tag{5}$$

Preliminary:

Extract textual information visually from Pdf utilizing Nougat (Neural Optical Understanding for Academic Documents)

Design: RoBERTa body with four different heads

Multitask-Learning Workflow:

- Knowledge Graph Embedding
- Detect entities from available classes (research problem, mathematical model, ...)
- Entity Linking (via scalar product)

pool is, the more complex it is to solve the line planning model in Phase 2. Hence, we require the number of lines in the line pool to be bounded by a value K to ensure tractability of Phase 2. To ensure that the line pool is 'good enough' is much harder. A minimal requirement for a good line pool is that it should at least be feasible for the line planning problem in Phase 2. Since there are so many different models for line planning it is unrealistic to guarantee feasibility for all possible models. However, the edge frequency constraints (1) are a minimal requirement which is used in most line planning models, and which we hence consider for designing a line pool.

We can now define the line pool generation problem for given K as follows.

(LPool) Given the PTN= (V, E) , lower and upper frequency bounds $f_e^{\min} \leq f_e^{\max}$ for all $e \in E$, and the set \mathcal{L}^0 of all cycle free paths in the PTN, find a set $\mathcal{L} \subseteq \mathcal{L}^0$ such that $|\mathcal{L}| \leq K$ and such that there exist $f_l \in \mathbb{N}_0$ for all $l \in \mathcal{L}$ with

$$f_e^{\min} \leq \sum_{\substack{l \in \mathcal{L} \\ e \in l}} f_l \leq f_e^{\max} \quad \forall e \in E.$$

Using binary variables

$$x_l = \begin{cases} 1 & \text{if line } l \text{ is chosen to be in the line pool} \\ 0 & \text{otherwise} \end{cases}$$

and variables $f_l \in \mathbb{N}_0$ to ensure the existence of a line concept, LPool can be formulated as the following integer program. To this end, let $M \geq \max_{e \in E} f_e^{\max}$.

$$\text{Find } x_l \tag{2}$$

$$\text{s.t. } \sum_{l \in \mathcal{L}^0} x_l \leq K \tag{3}$$

$$\sum_{\substack{l \in \mathcal{L}^0 \\ e \in l}} f_l \geq f_e^{\min} \quad \forall e \in E \tag{4}$$

$$\sum_{\substack{l \in \mathcal{L}^0 \\ e \in l}} f_l \leq f_e^{\max} \quad \forall e \in E \tag{5}$$

Preliminary:

Extract textual information visually from Pdf utilizing Nougat (Neural Optical Understanding for Academic Documents)

Design: RoBERTa body with four different heads

Multitask-Learning Workflow:

- Knowledge Graph Embedding
- Detect entities from available classes (research problem, mathematical model, ...)
- Entity Linking (via scalar product)
- Relation Extraction

number represents the maximum number of vehicles that are allowed to traverse each edge and may model capacity or security restrictions. For a line concept with frequencies $f_l \in \mathbb{N}_0$ for all $l \in \mathcal{L}$ it is then required that

$$f_e^{\min} \leq \sum_{\substack{l \in \mathcal{L}: \\ e \in l}} f_l \leq f_e^{\max} \quad \forall e \in E. \quad (1)$$

in defining
formulation /
contains

defining
formulation

line planning it is unrealistic to guarantee feasibility for all possible models. However, the edge frequency constraints (1) are a minimal requirement which is used in most line planning models, and which we hence consider for designing a line pool.

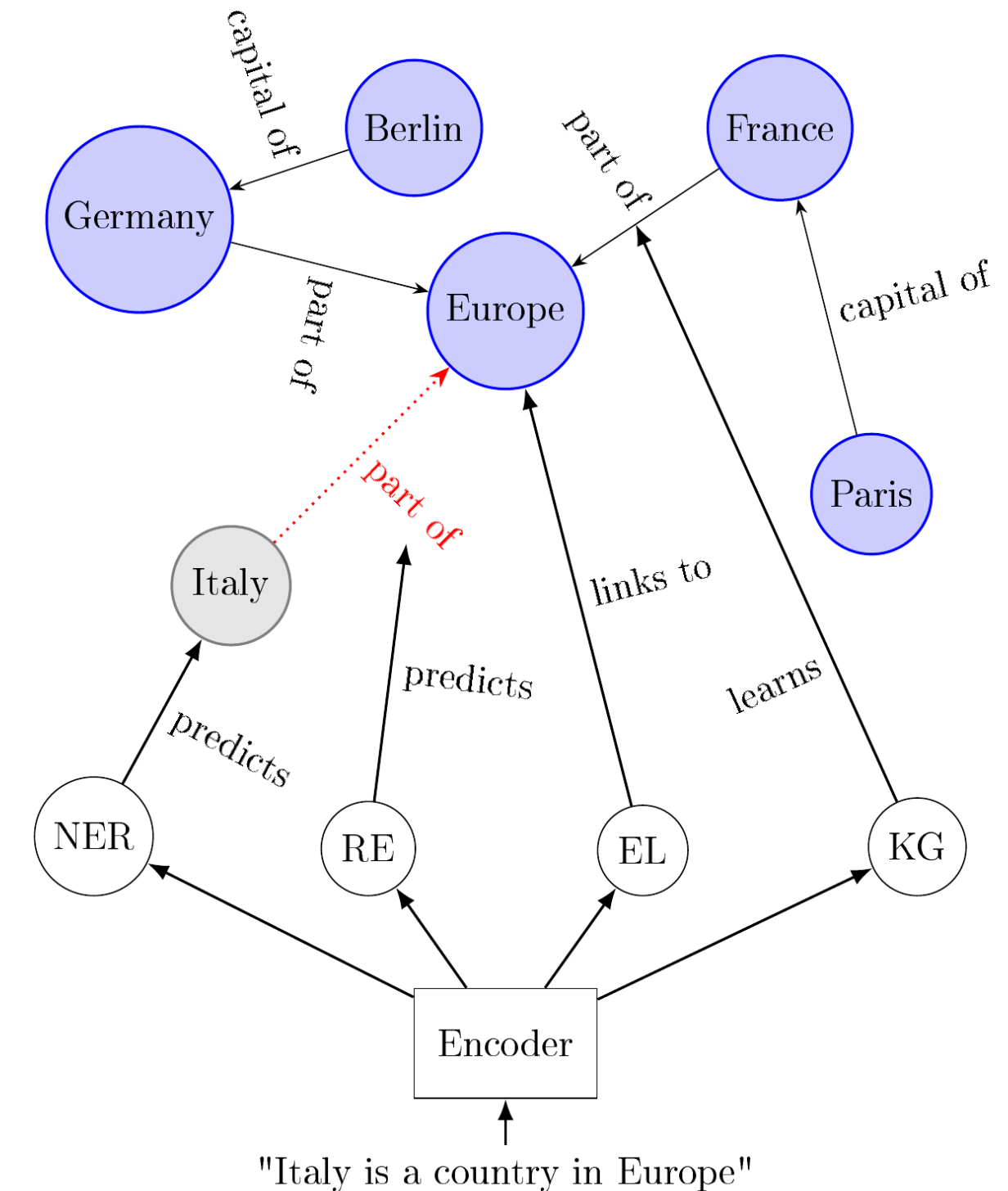
Preliminary:

Extract textual information visually from Pdf utilizing Nougat (Neural Optical Understanding for Academic Documents)

Design: RoBERTa body with four different heads

Multitask-Learning Workflow:

- Knowledge Graph Embedding
- Detect entities from available classes (research problem, mathematical model, ...)
- Entity Linking (via scalar product)
- Relation Extraction



So far:

- No complete workflow → no final results
- Training and finetuning of models require a fair amount of complex data: Publications with annotated entities, relations and links
- Some parts work well (NER), some parts are really difficult (relation extraction due to spurious relations)

Future plans:

- Build on workflow and finetune models with more data
- Implement “human in the loop” approach
- Utilize MaRDMO chatbot → chatbot generates proposal of entities and their relations, human can validate them (requires decoder)