ORKGEx: Leveraging Language and Vision Models with Knowledge Graphs for Research Contribution Annotation

Hassan Hussein , Fahad Ahmed , Allard Oelen , Ralph Ewerth , Sören Auer

Hassan Hussein

TIB – Leibniz Information Centre for Science and Technology

IARIA- Web 2025, Lisbon
13, March, 2025

Bibliography

1. **Computer Information System**
   Tompkins Cortland Community College Tompkins Cortland Community College, NY 2009

2. **Interne at the legal institute in Web Development Area**
   Cornell University, NY 2009

3. **Web Applications Development Team Leader**
   Telecom Egypt 2011 to 2017

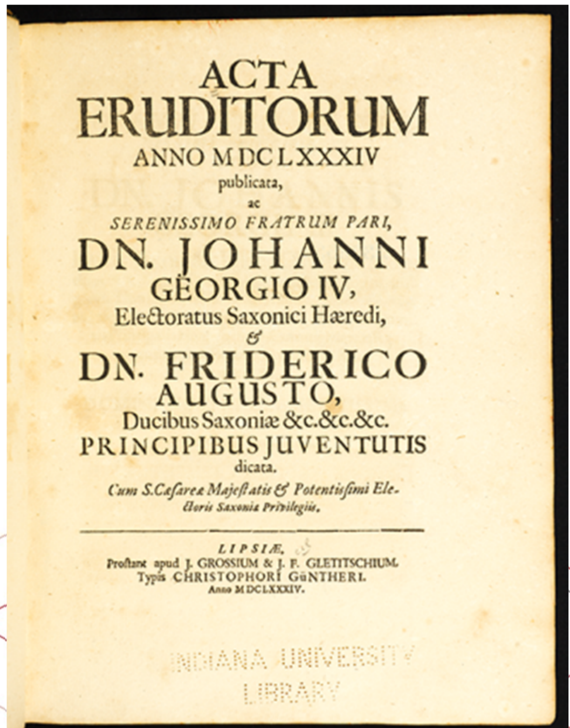
4. **M.Sc. in Human Computer Interaction**
   Siegen University, Germany 2021

5. **PhD Student in User Interface for Semantic Web**
   Leibniz University, Germany

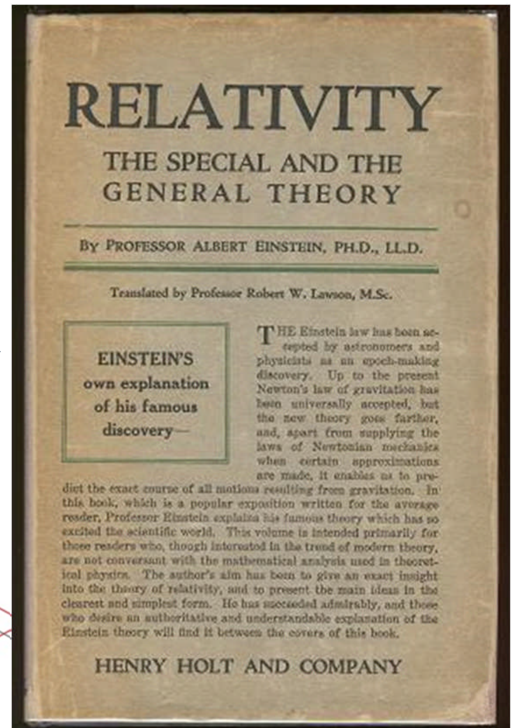
1. Introduction

Scholarly Communication Over Time

Over 300 years ago


ACTA
ERUDITORUM
ANNO MDCLXXXIV
publicata,
ac
SERENISSIMO FRATRUM PARI,
DN. JOHANNI
GEORGIO IV,
Electoratus Saxonici Hæredi,
&
DN. FRIDERICO
AUGUSTO,
Ducibus Saxoniæ &c.&c.&c.
PRINCIPIBUS JUVENTUTIS
dicata.
Cum S.Cæsareæ Majestatis & Potentissimi Ele-
ctoris Saxoniæ Privilegiis.
LIPSIÆ,
Prostant apud J. GROSSIUM & J. F. GLETITSCHIUM.
Typis CHRISTOPHORI GÜNTHERI.
Anno MDCLXXXIV.

100 years ago


RELATIVITY
THE SPECIAL AND THE
GENERAL THEORY
BY PROFESSOR ALBERT EINSTEIN, PH.D., LL.D.
Translated by Professor Robert W. Lawson, M.Sc.
EINSTEIN'S own explanation of his famous discovery—
THE Einstein law has been accepted by astronomers and physicists as an epoch-making discovery. Up to the present Newton's law of gravitation has been universally accepted, but the new theory goes farther, and, apart from supplying the laws of Newtonian mechanics when certain approximations are made, it enables us to predict the exact course of all motions resulting from gravitation. In this book, which is a popular exposition written for the average reader, Professor Einstein explains his famous theory which has so excited the scientific world. This volume is intended primarily for those readers who, though interested in the trend of modern theory, are not conversant with the mathematical analysis used in theoretical physics. The author's aim has been to give an exact insight into the theory of relativity, and to present the main ideas in the clearest and simplest form. He has succeeded admirably, and those who desire an authoritative and understandable explanation of the Einstein theory will find it between the covers of this book.
HENRY HOLT AND COMPANY

20 years ago


Information Retrieval — P. BAXENDALE, Editor

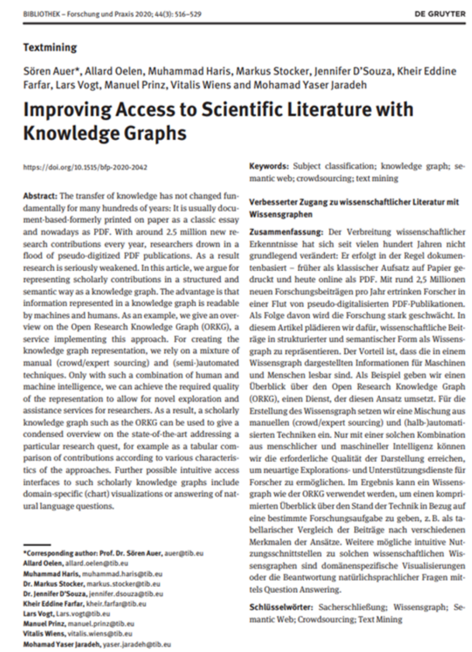
A Relational Model of Data for Large Shared Data Banks

E. F. Codd
IBM Research Laboratory, San Jose, California

1. Relational Model and Normal Form

Volume 13 / Number 6 / June, 1970 — Communications of the ACM 377

Today


BIBLIOTHEK – Forschung und Praxis 2020; 44(3): 516–529 — DE GRUYTER

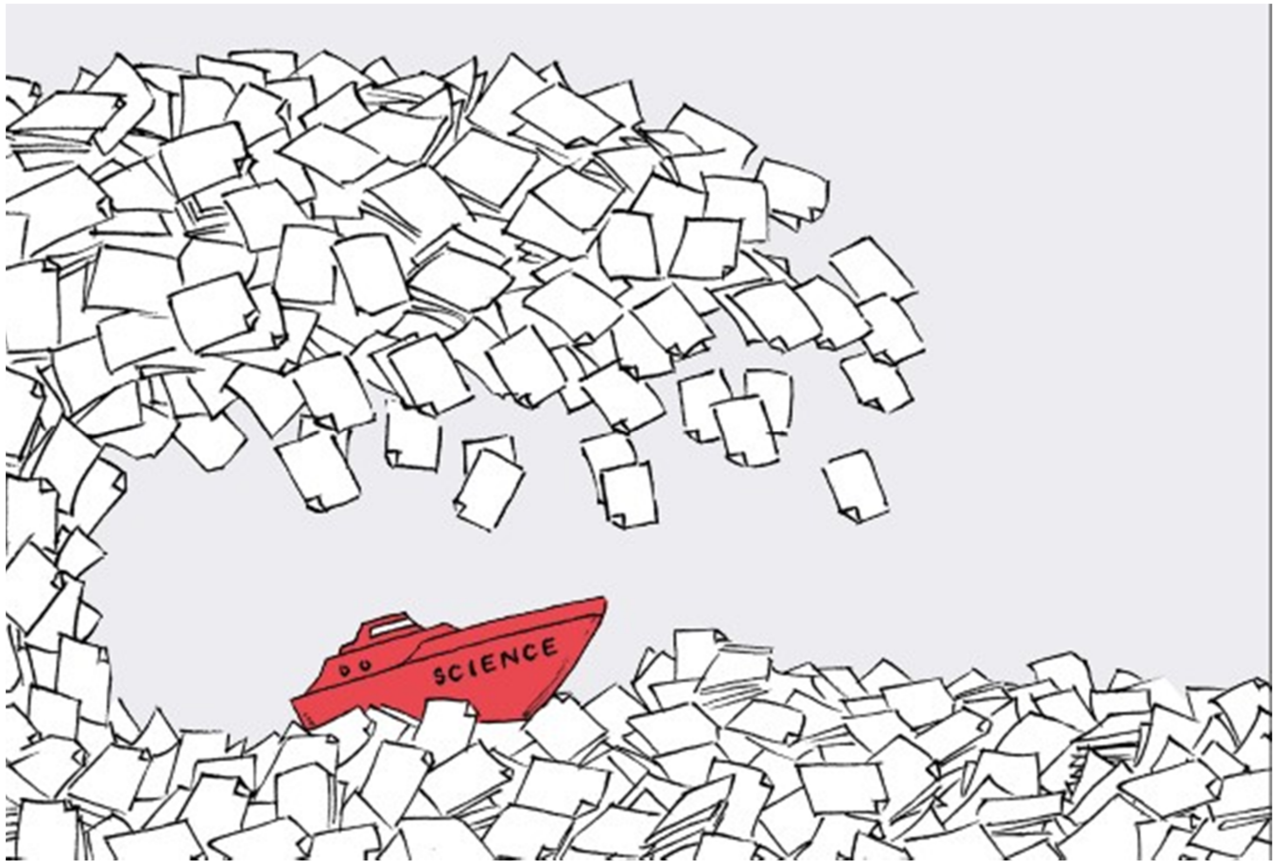
Textmining

Sören Auer*, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer D'Souza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens and Mohamad Yaser Jaradeh

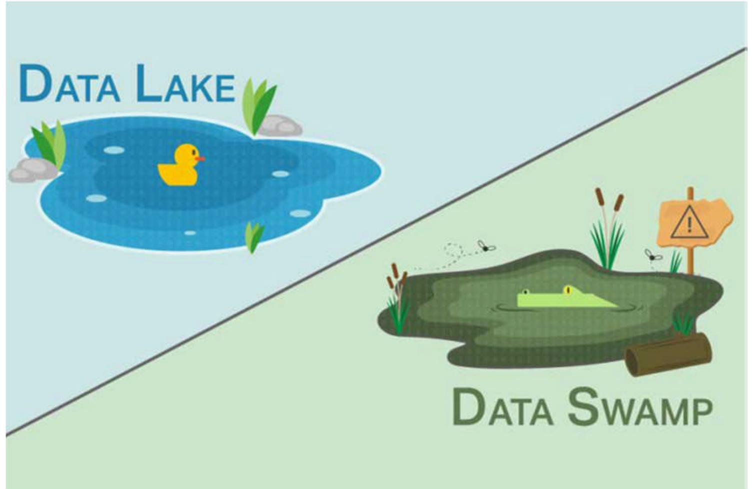
Improving Access to Scientific Literature with Knowledge Graphs

Reproducibility Crisis

1. ~ 2.5 million new publications per year
2. Globally ~ $1.7 trillion spent on research
3. Monopolization of commercial actors
4. Deficiency of Peer-Review
5. Predatory Publishing



The Data Swamp Problem

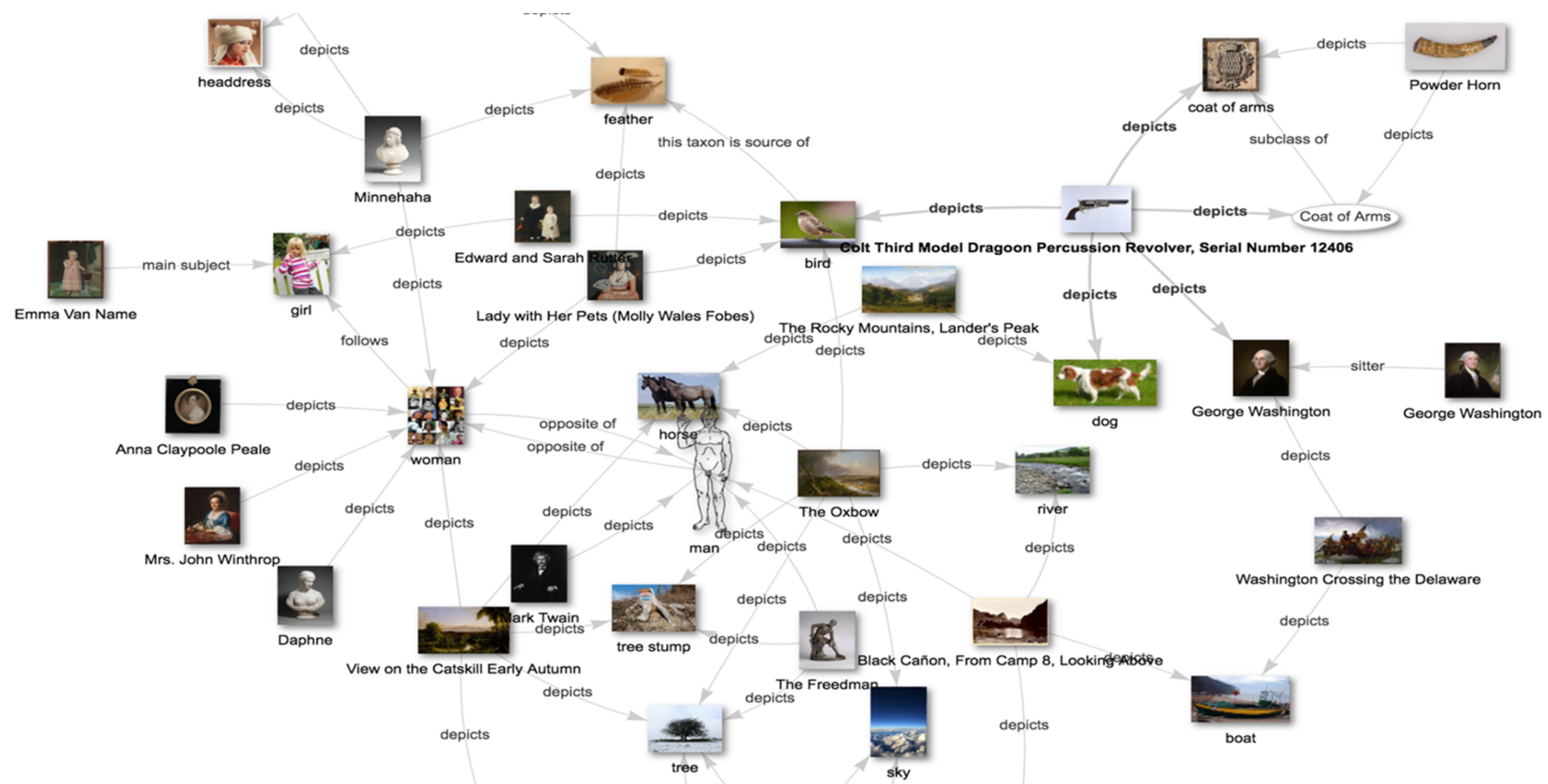


1. **Semantic Description of Research Contributions**
   A. Researchers often struggle to clearly and accessibly convey their work.
   B. The annotation process is time consuming and cumbersome
2. **Information Overload**
   A. The exponential growth of scientific publications has led to information overload
   B. Insufficiently automatized and lack user integration
3. **Lack of Standardization**
   A. There is a lack of standardization in how research contributions are annotated and shared.
   B. This inconsistency fragments knowledge and hinders effective meta-analyses.

2. Lighthouse in the Flood

Knowledge Graphs (KGs)

Knowledge Graphs (KGs): Data structures that represent knowledge in a graph format, where nodes represent entities (e.g., people, places, things) and edges represent relationships between these entities.



Source: Wikidata

Symbolic Representation: KGs use formal knowledge representation languages (such as RDF and OWL) to encode facts, making the information machine-readable and understandable.

KGs: Accuracy & Integrity

1. Accuracy & Reliability

A. Designed to store and retrieve factual information with high accuracy.

2. Structured Data

A. Uses a graph structure for efficient organization and retrieval of interrelated data.

3. Knowledge Graphs have Factual Information

A. **Enhanced Search and Querying:** Enables more precise searches by understanding relationships.

B. **Knowledge Discovery:** Links related information to uncover new insights.

C. **Contextual Understanding:** Provides broader context for better comprehension.

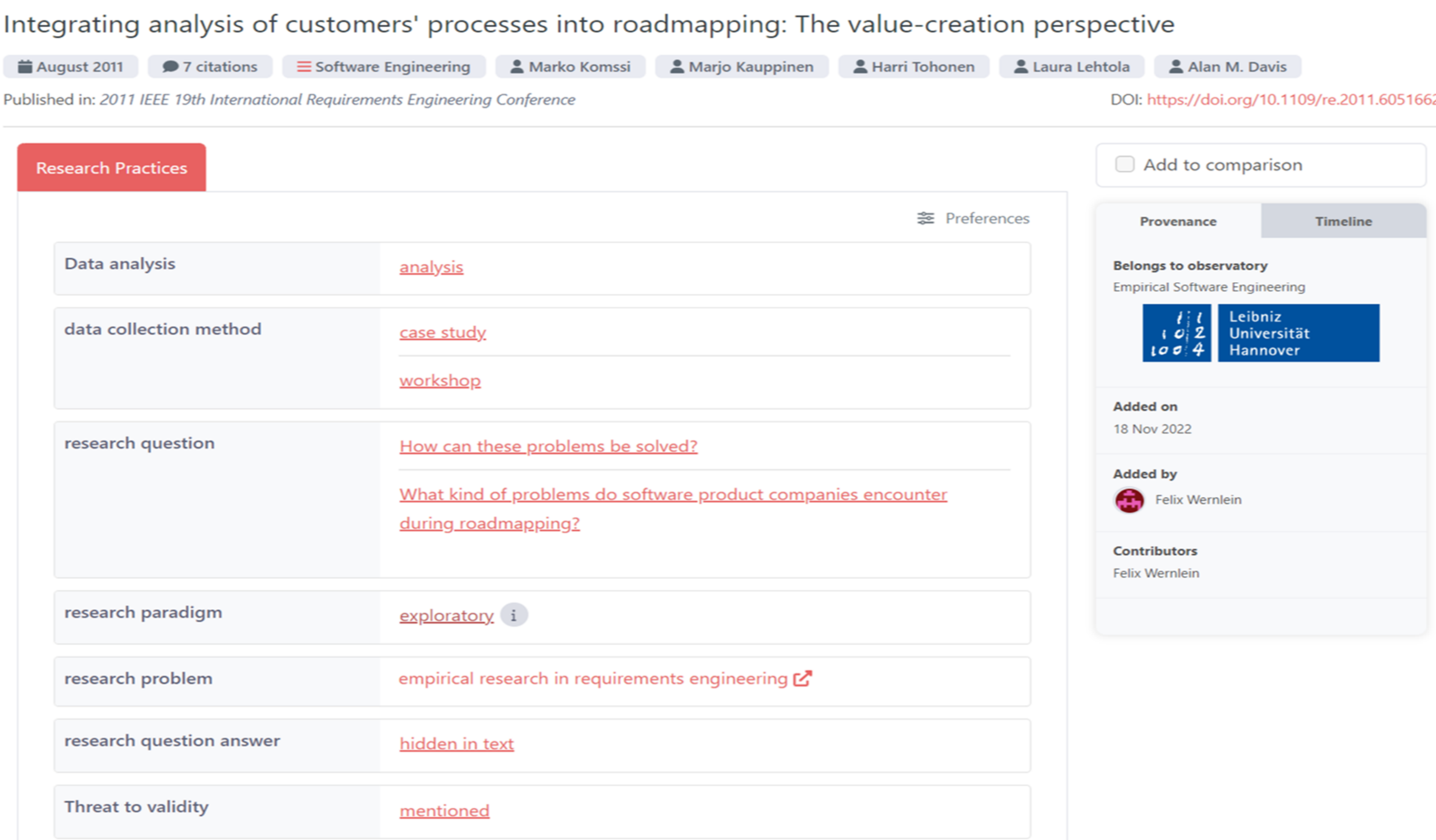
D. **Symbolic Representation:** Uses symbols and formal languages to represent structured knowledge

E. **Human and Machine Interpretability:** Ensures both humans and machines can process and understand the data.

F. **Interoperability:** Supports data integration using standardized formats

Open Research Knowledge Graph (ORKG)

ORKG- Paper View



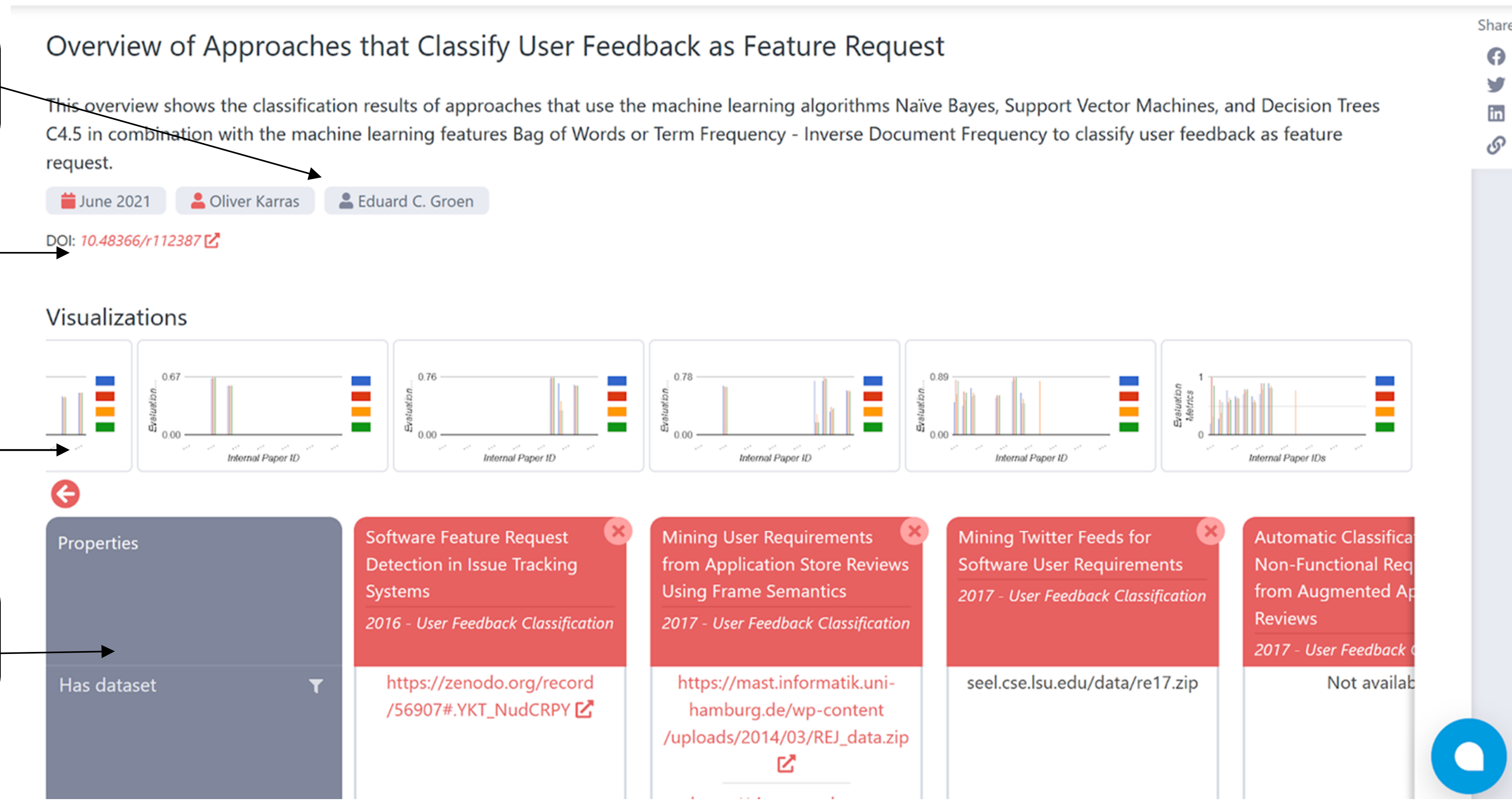
Publishing State-of-the-Art comparisons

Acknowledgement of creators

Citable DOI

Created visualizations

Interactive comparison with filtering



https://orkg.org/comparison/R112387/

3. Problem

ORKG Helps but Challenges Remain

1. Extensive Editing Required
2. Property Selection Struggles
3. Semantic Descriptions Hard to Find
4. Time-Consuming Annotation
5. Limited Paper Representation (only text)
6. Lack of Motivation

The Rise of Generative AI

1. Transforming Knowledge Processing

A. AI models can automate knowledge extraction

B. Promise of reducing manual effort in annotation

C. Potential for better semantic understanding

BUT…. new challenges emerged!

Challenges with Neural Models

1. **Non-Deterministic Behavior**
   A. LLMs can generate different outputs for the same input, making consistency a challenge
2. **Opaqueness (Lack of Transparency)**
   A. Users cannot see or understand how decisions are made.
3. **Trust and Adoption Issues:**
   A. Users may hesitate to trust AI systems they don't understand
4. **Tendency to Confabulate:**
   A. LLMs can generate plausible but incorrect or nonsensical information ("hallucinations").
   B. Persistent errors reduce confidence.
   C. Risk of spreading false information.

4. Goal

Minimal invasive Interaction

1. We can achieve that by:

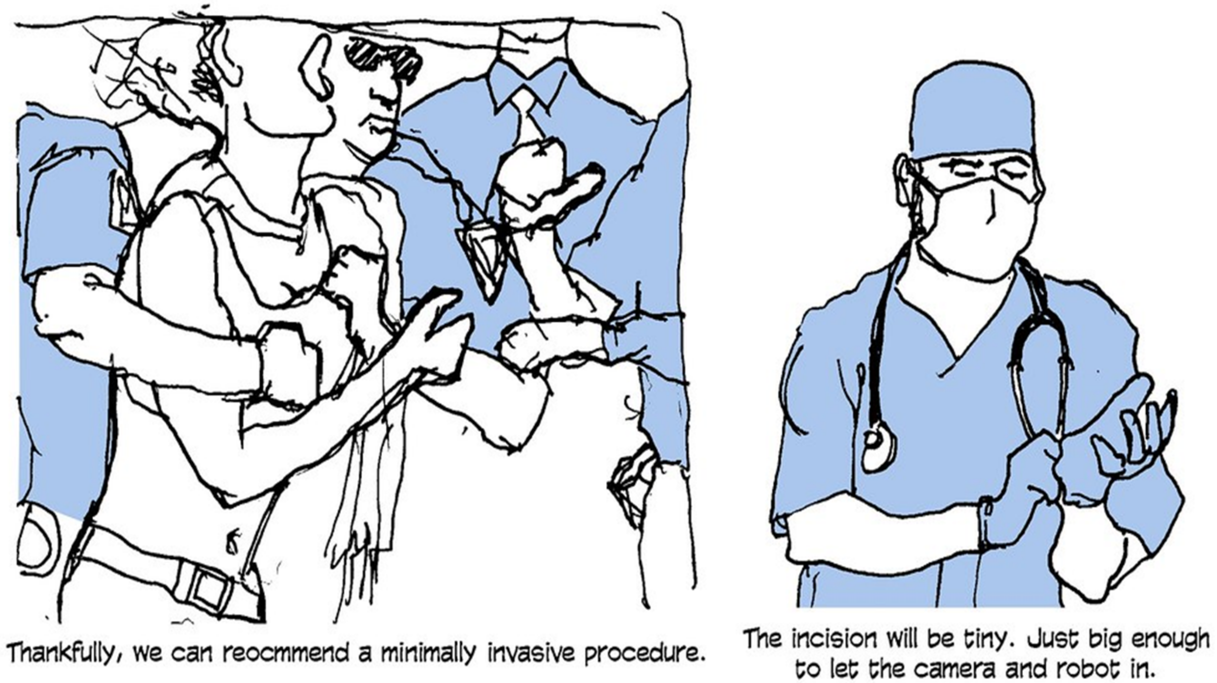
A. Integrating the annotation process into the researchers ecosystem

B. Automatically extract Metadata and relevant information

2. Leverageing the AI Techniques to:

A. Minimize humans efforts

B. Save the time



Soure: Greg Borenstein

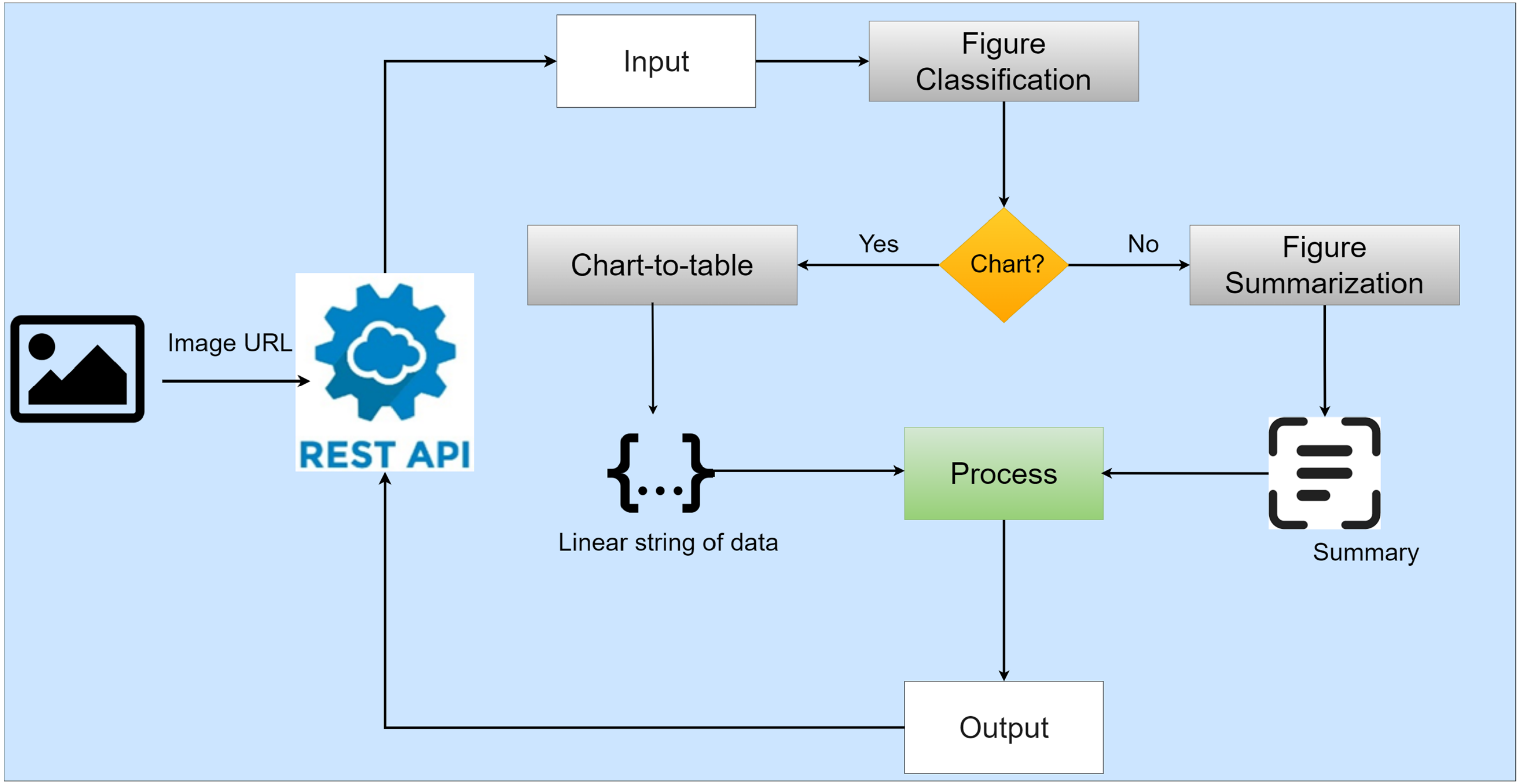
3. Seamless Integration to:

A. Harness both human intelligence and advanced neural and symbolic AI techniques

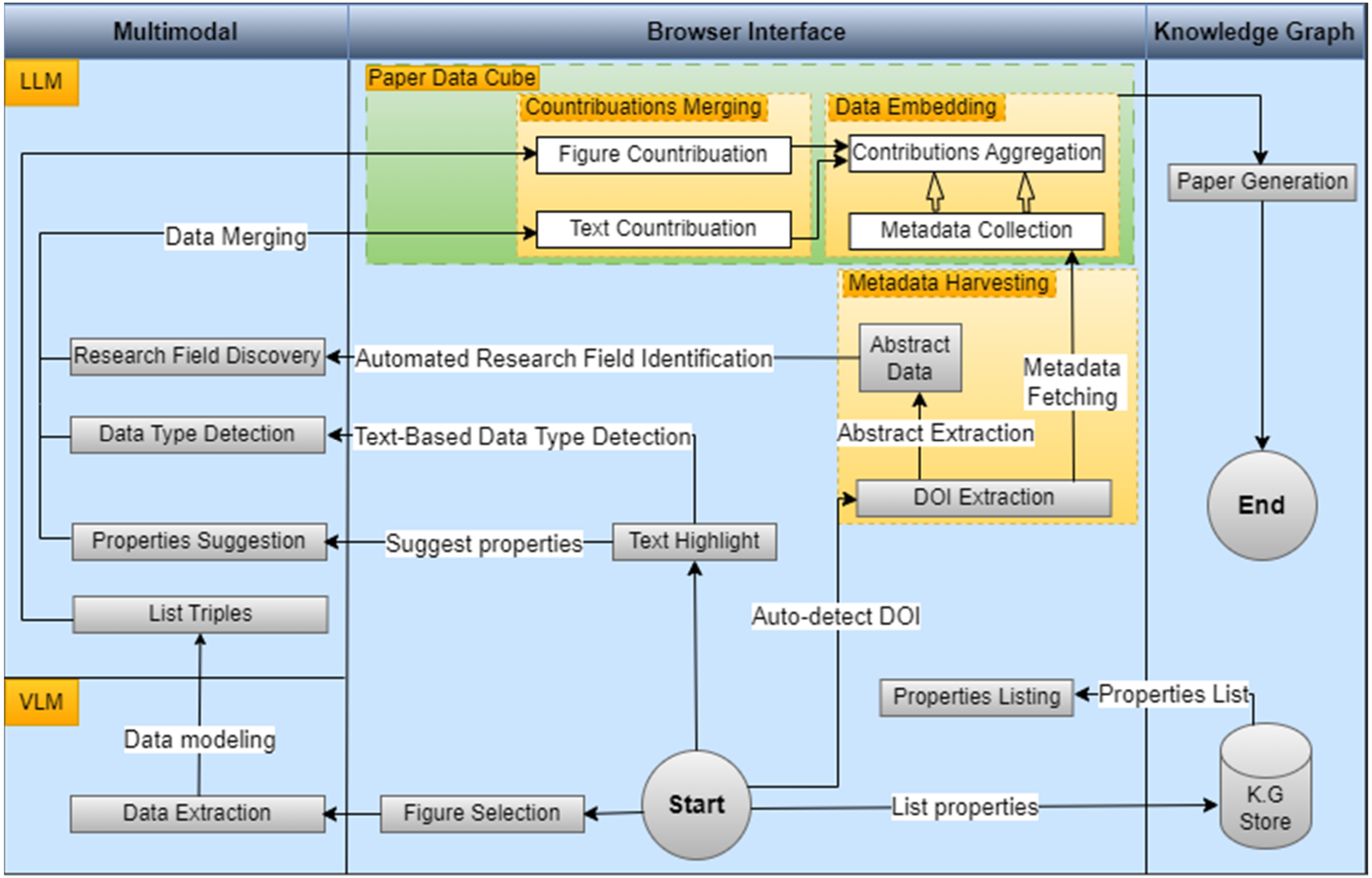
B. Integrate user contributions in a structured manner

5. Approach

Figure Data Extraction Pipeline



Human-Centric Multi-Modal AI Annotation Pipeline



5. Evaluation

User Study and Evaluation

1. **Participants:** 11 professionals (Postdocs, PhDs, Developers)

2. **82% had prior ORKG experience:** Provided informed feedback

3. **Key Findings:**

Feature	Mean ± SD	95% CI	Key Insight
Figure Triples Extraction	4.27 ± 0.65	(3.83, 4.71)	Highly efficient
Property Suggestions	3.73 ± 1.14	(2.97, 4.49)	Variable performance
Overall Speed	4.82 ± 0.39	(4.56, 5.08)	Significant gain

4. **Reliability and Performance Analysis:**

Metric	Mean ± SD	Distribution	95% CI
Data Type Detection	4.09 ± 0.67	5★: 27.3%, 4★: 54.5%, 3★: 18.2%	(3.64, 4.54)
Research Field Classification	4.00 ± 0.74	5★: 27.3%, 4★: 45.5%, 3★: 27.3%	(3.50, 4.50)
Metadata Extraction	4.27 ± 0.86	5★: 54.5%, 4★: 18.2%, 3★: 27.3%	(3.69, 4.85)

User Study and Evaluation

5. User Feedback Highlights

A. **82%** reported **faster annotation** than traditional interfaces

B. **90% effectiveness** in **figure-based triple extraction**

C. **82% trust** in AI-generated content, with **73% accuracy perception**

D. **High satisfaction (Mean: 4.18 ± 0.75)**

E. **Strong AI-assisted performance**: metadata extraction (4.27 ± 0.86), data type detection (4.09 ± 0.67)

6. Areas for Improvement

A. AI-generated **property suggestions** need more **contextual awareness**

B. Tooltips should have **better visibility** (larger fonts, noticeable colors)

6. Limitaion

Limitations Identified:

1. Performance Measurement Challenges:

A. User-reported speed improvements (4.82 ± 0.39), but no absolute baselines due to

1. Annotator expertise and familiarity

2. Paper complexity (length, structure, content)

3 .Number and complexity of figures

2. System Evaluation Challenges:

A. The need for Larger-scale comparison (manual vs. automated annotations)

B. Controlled environment for quantitative time measurements

C. Standardized test sets with varying complexity

3. Sample Diversity:

A. Evaluation with 11 participants (82% with ORKG experience)

B. Small sample size limits generalizability, requiring broader validation with diverse expertise levels and larger samples

7. Future Work

Future Directions:

1. Explore advanced techniques in computer vision and NLP (e.g., transformer-based models for figure extraction).

2. Address **scalability challenges** for large documents and multi-user annotation synchronization.

3. Investigate **deployment challenges** (browser version consistency, complex figure processing).

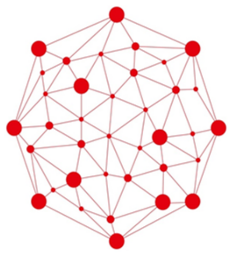
4. Expand **evaluation scope** with larger, diverse research communities to mitigate biases (academic disciplines, experience levels).

5. Enhance **support for complex data types** (interactive tables) and diverse use cases.

8. Demo

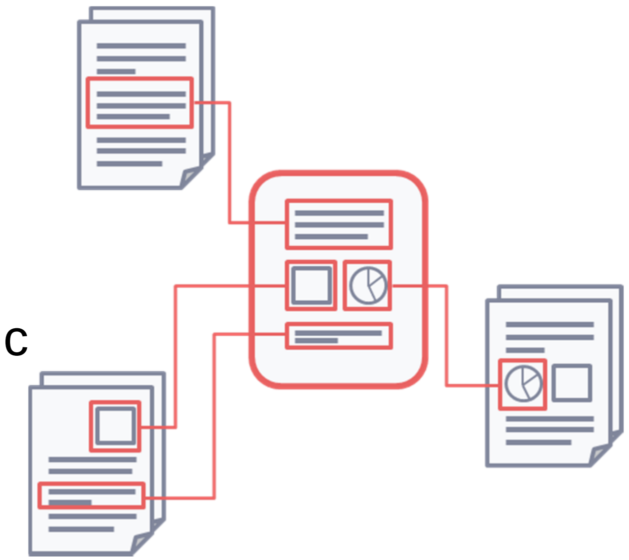
Key Takeaways

Human-AI synergy enhances research capabilities: Curation workflows with machine assistance (LLM and VLM) and human-in-the-loop refine knowledge representations in the ORKG.



Clarity and consistency of Knowledge Representation: Knowledge graph encapsulates factual information in a symbolic form that is accessible to both humans and machines.

Collaborative framework fosters reproducibility: Transparent and accurate knowledge representation through crowd work in the KG makes it easier for other researchers to verify and reproduce study results iteratively.



Questions