



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



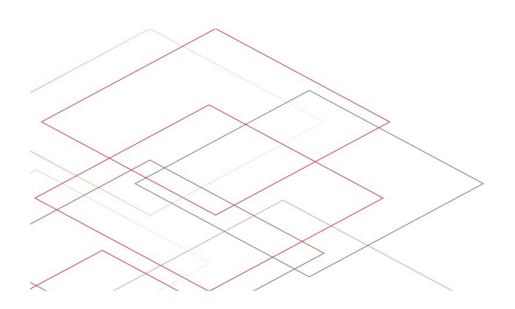








1. Introduction





LEIBNIZ INFORMATION CENTRE

DE GRUYTER

Scholarly Communication Over Time

Over 300 years ago

ACTA

FRUDITORUM

ANNO M DC L XXXIV

publicata,

SERENISSIMO FRATRUM PARI,

DN. JOHANNI

GÉORGIO IV,

Electoratus Saxonici Hæredi,

DN. FRIDERICO AUGUSTO,

Ducibus Saxoniæ &c.&c.&c.

PRINCIPIBUS JUVENTUTIS

dicata

Cum S.Cafarea Majestatis & Potentissimi Ele.

Eloris Saxonia Privilegiis.

LIPSIA.

Profiant apud J. GROSSIUM & J. F. GLETITSCHIUM. Typis CHRISTOPHORI GüNTHERL Anno MDCLXXXIV.

100 years ago

THE SPECIAL AND THE

GENERAL THEORY

BY PROFESSOR ALBERT EINSTEIN, PH.D., LL.D.

Translated by Professor Robert W. Lawson, M.Sc.

liet the exact course of all motions resulting freen gravitation. In

this book, which is a popular expection written for the average reader, Professor Rinstein explains his famous theory which has no

excited the scientific world. This volume is intended primarily for

those readers who, thoogh interested in the trend of modern theory.

re not conversant with the mathematical analysis used in theoret-

cal physics. The author's aim hus been to give an exact insight

into the theory of relativity, and to present the main ideas in the

learest and simplest form. He has succeeded admirably, and these who desire an authoritative and understandable explanation of the instein theory will find it between the covers of this book.

HENRY HOLT AND COMPANY

THE Einstein law has been ac-

tepted by astronomers and physicists as an epoch-making

discovery. Up to the present

Newton's law of gravitation has

here universally accepted, but

the new theory goes farther,

and, apart from supplying the

laws of Newtonian mechanics when cortain approximations

are made, it enables us to pre-

RELATI

EINSTEIN'S

own explanation

of his famous

discovery-

20 years ago

Information Retrieval

A Relational Model of Data for

Rutre usen of large data banks must be protected from having to know how the data is angualized in the machine file element preperentation). A prospilar service which supplies such alformation in and a splitationary paydown, Archivies of users at terminolis and next applications programs should remain undifficted when the internal representation of data is donged and serve when some acapted of the external representation and serve when some acapted of the external representation of a serve when some acapted of the external representation and serve when some acapted of the external representation and serve when some acapted of the external representation and serve when some acapted of the external representation and serve when some acapted of the external representation and serve when some acapted of the external representation and serve when some acapted of the external representation and serve when serve and the external representation and serve when serve and the server and the

and even when some oppers of the external representation are changed. Changes in data: representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information. Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network

models of the data. In Section 1, inadequacies of these models

modes or me dord, in Section 1, modequotes or mese modes are discussed. A model based on n-ory relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain opera-tions on relations (other than logical inference) are discussed

and applied to the problems of redundancy and consistency

EEY WORDS AND PHRASES data bank, data base, data structure, data argunization, hiarantika ed data, naturuka of data, relationa, derivability, redundancy, compatibiliton, join, relativa llanguage, predicate calkula, security, data lategrity CF CATFOORES 370, 373, 375, 420, 422, 429

This paper is concerned with the application of ele-

This paper is concerned with the application of ele-mentary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Levein and Maron [2] provide numerous references to work

In contrast, the problems treated here are those of data

an contrast, the proteins treated nere are those of dala independence—the independence of application programs and terminal activities from growth in data types and changes in data representation—and certain kinds of data inconsistency which are expected to become troublesome

1. Relational Model and Normal Form

in the user's model.

1.1. INTRODUCTION

even in nondeductive syst Volume 13 / Number 6 / June, 1970

in this area.

Large Shared Data Banks

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

BIBLIOTHEK - Forschung und Praxis 2020; 44(3): 516-529

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

Todav

Improving Access to Scientific Literature with **Knowledge Graphs**

https://doi.org/10.1515/bfp-2020-2042 Keywords: Subject classification; knowledge graph; se

mantic web: crowdsourcing: text mining Abstract: The transfer of knowledge has not changed fun-damentally for many hundreds of years: It is usually docu-wissensgraphen wissensgraphen erter Zugang zu wis

ment-based-formerly printed on paper as a classic essay and nowadays as PDF. With around 2.5 million new re- Zusammenfassung: Der Verbreitung wisse and however, the second semantic way as a knowledge graph. The advantage is that neuen Forschungsbeiträgen pro Jahr ertrinken Forscher in semantic way as a knowledge graph. Ine advantage is that never roticnumgboritagen pro junt ertmixen roticiter in information represented in a knowledge graph is reachable ere Flur von piecelo digitalisiente POP-Probibiliationen. by machines and humans. Ka na example, we give a nover – Ah Folge davon wird die Forschung stark geschwicht. In view on the Open Research Knowledge Graph (ORKG), a diesem Artikel plädieren wir dafür, wissenschaftliche Beit-reice Implementing this approach. For exeting the single instrukturieret und semantischer Form als Wissens-treice Implementing this approach. For exeting the knowledge graph representation, we rely on a mixture of graph zu repräsentieren. Der Vorteil ist, dass die in einem manual (crowd/expert sourcing) and (semi-lautomated Wissensgraph dargestellten Informationen für Maschinen manual (covovecpent sourcing) and beni-jauronanes waverngapio usgestenes mormanismonth in associates techniques. Only with such a combination of human and un Menschne Hesbari, dir. Als Biejseige deres wir einen machten Intelligence, we can achieve the required quality. Überblick über den Open Research Knowledge Graph of the representation to allow for novel-sequencing and (ORKG), einen Dienst, der diesen Anatz unset. Für die assistance services for researchers, As a result, a scholarly. Exstellung des Wissensgraph setzen wir eine Mischung aus knowledge graph such as the ORKG can be used to give a manuellen (crowd/expert sourcing) und (halb-)automati condensed overview on the state-of-the-art addressing a particular research quest, for example as a tabular com aution of combutos according to various characteristic wirt die erfortleriche Qualität der Darstellung erreichen, tics of the approaches. Further possible intuitive access um neuartige Explorations- und Unterstützungsdienste für interfaces to such scholarly knowledge graphs include Forscher zu ermöglichen. Im Ergebnis kann ein Wissens domain-specific (chart) visualizations or answering of nat- graph wie der ORKG verwendet werden, um einen kompr gaapi wie die Orko verwender werden, uie einen kompi-mierten Überblick über den Stand der Technik in Bezug auf eine bestimmte Forschungsaufgabe zu geben, z. B. als ta-bellarischer Vergleich der Beiträge nach verschiedenen Merkmalen der Ansätze. Weitere mögliche intuitive Nutzungsschnittstellen zu solchen wissenschaftlichen Wisagraphen sind domänenspezifische Visualisierunger er die Beantwortung natürlichsprachlicher Fragen mit

al Prinz, manuel.prinz@tib.eu talis Wiens, vitalis.wiens@tib.eu

tels Question Answering. Schlüsselwörter: Sacherschließung: Wissensgraph: Semantic Web: Crowdsourcing: Text Mining

iral language questions Carresponding author: Prof. Dr. Sören Auer, auer@tib.eu Allard Oelen, allard.oelen@tib.eu Muhammad Naris, muhammad.haris@tib.eu Dr. Markas Söreker, markus.stockrefitib.eu Dr. Jennifer O'Souza, Jennifer.diouza@tib.eu Huhair Kalim särjark. Johie Lafardib.eu







P. BAXENDALE, Edito

The relational view (or model) of data described in

Ine relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for non-inferential systems. It provides a means of describing data with its natural structure only—that is, without superim-

posing any additional structure for machine representatio

data ianguage which will yied maximal independence be-tween programs on the one hand and machine represent-tion and organization of data on the other. A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy,

and consistency of relations-these are discussed in Section

and consistency of relations—these are discussed in Section 2. The network model, on the other hand, has paramed a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of rela-tions (see remarks in Section 2 on the "connection trap"). Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted

data systems, and also the relative merits (from a logical

standpoint) of competing representations of data within a single system. Examples of this clearer perspective are cited in various parts of this paper. Implementations of

systems to support the relational model are not discussed

data language which will yield maximal inde

osses. Accordingly, it provides a basis for a high level

elements to be stored in at least one total ordering which is closely associated with the hardware-determined ordering of addresses. For example, the records of a file concerning parts might be stored in ascending order by part serial number. Such systems normally permit application programs to assume that the order of presentation of record from such a file is identical to (or is a subordering of) the

Communications of the ACM

mad Yaser Jaradeh, yaser.jarad

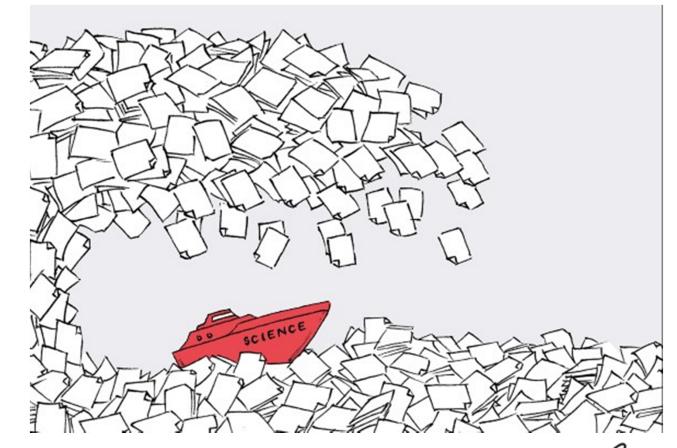
IB FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY



IB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

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

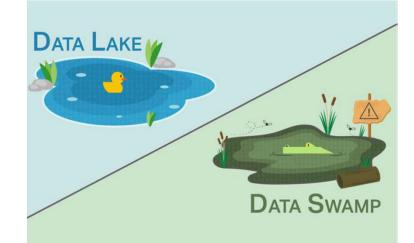
2. Information Overload

- A. The exponential growth of scientific publications has led to information overload
- B. Insufficiently automatized and lack user integration

3. Lack of Standardization

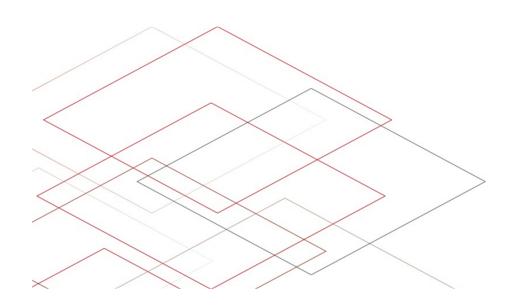
- A. There is a lack of standardization in how research contributions are annotated and shared.
- B. This inconsistency fragments knowledge and hinders effective meta-analyses.







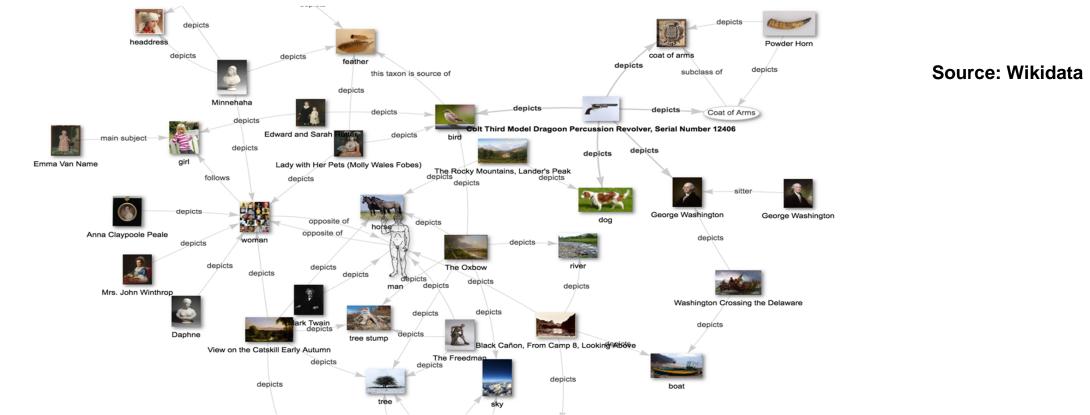
2. Lighthouse in the Flood





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.



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)







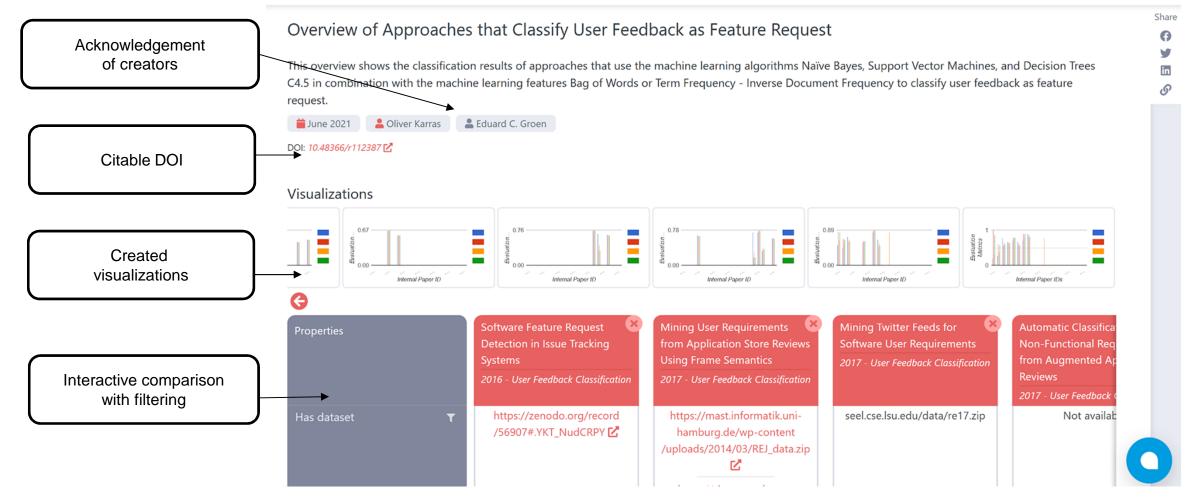
ORKG- Paper View

Integrating analysis of customers' processes into roadmapping: The value-creation perspective

arch Practices		Add to comparison	
	🐲 Preferences	Provenance Timeline	
Data analysis	analysis	Belongs to observatory Empirical Software Engineering	
lata collection method	<u>case study</u>	lil Leibniz lo2 Universität Lo34 Hannover	
	workshop	Added on	
esearch question	How can these problems be solved?	18 Nov 2022	
	What kind of problems do software product companies encounter during roadmapping?	Added by Felix Wernlein	
		Contributors Felix Wernlein	
esearch paradigm	exploratory i		
esearch problem	empirical research in requirements engineering 🖸		
esearch question answer	hidden in text		

Associat

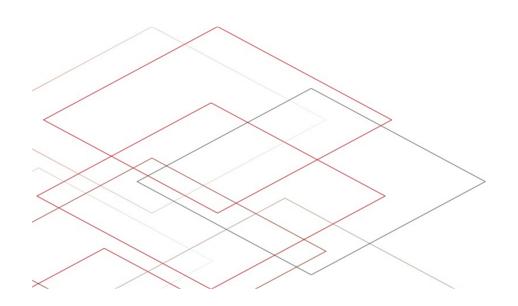
Publishing State-of-the-Art comparisons



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



3. Problem

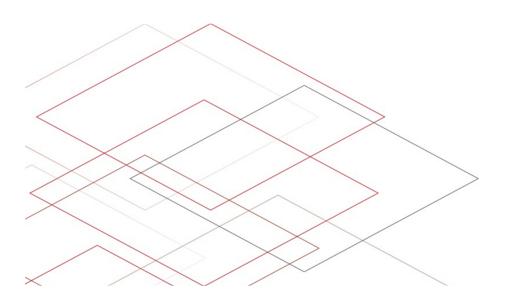






ORKG Helps but Challenges Remain

- 1. Extensive Editing Required
- 3. Semantic Descriptions Hard to Find
- 5. Limited Paper Representation (only text)



- 2. Property Selection Struggles
- 4. Time-Consuming Annotation
- 6. Lack of Motivation

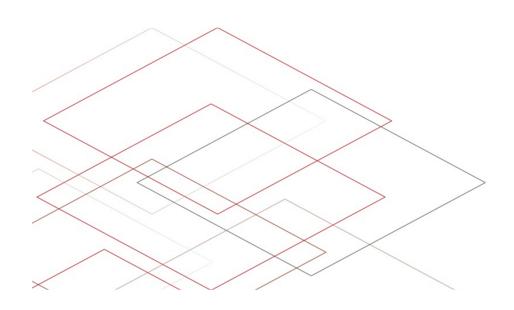




The Rise of Generative Al

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:

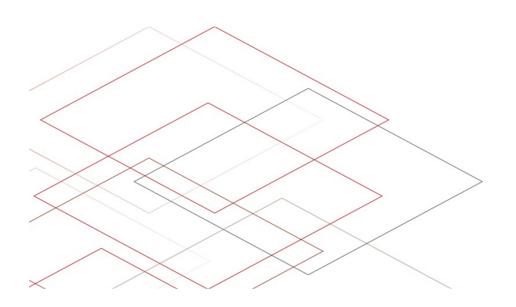
A. LLMs can generate plausible but incorrect or nonsensical information ("hallucinations").

B. Persistent errors reduce confidence.

C. Risk of spreading false information.



4. Goal





Minimal invasive Interaction



LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

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
- 3. Seamless Integration to:

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

B. Integrate user contributions in a structured manner





Thankfully, we can recommend a minimally invasive procedure.

Soure: Greg Borenstein

to let the camera and robot in.





5. Approach

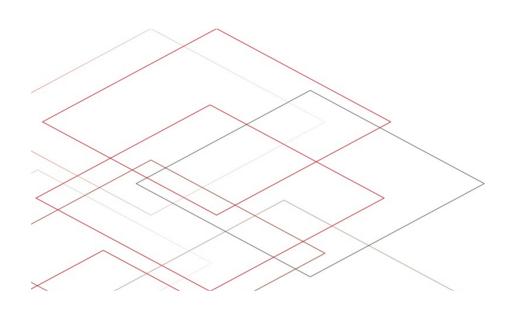
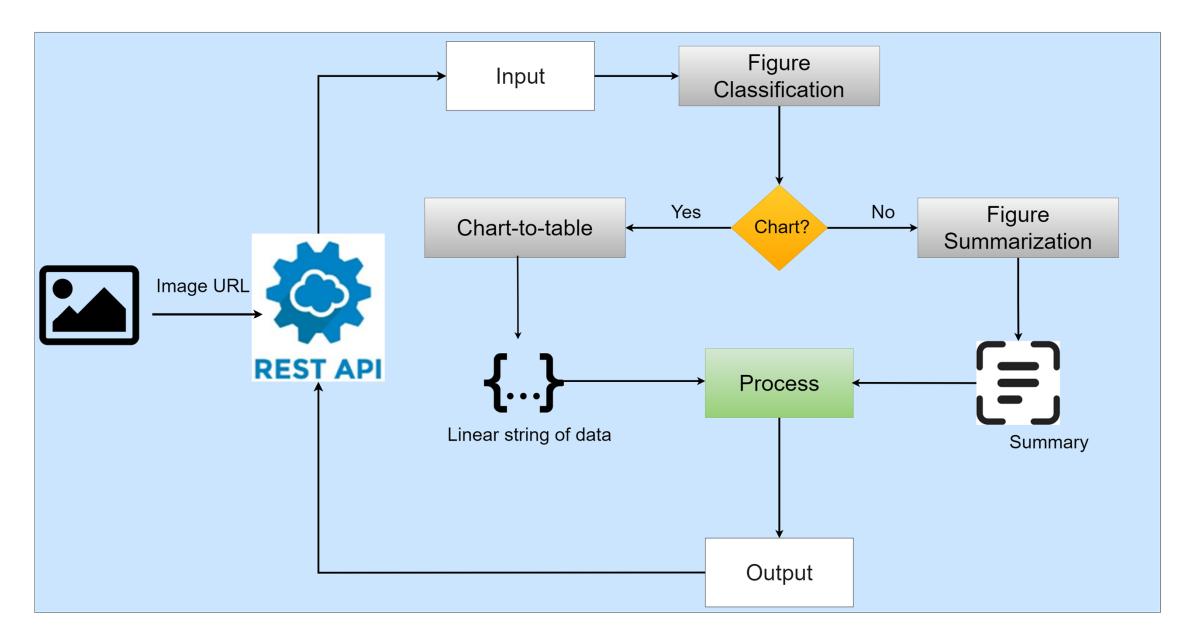
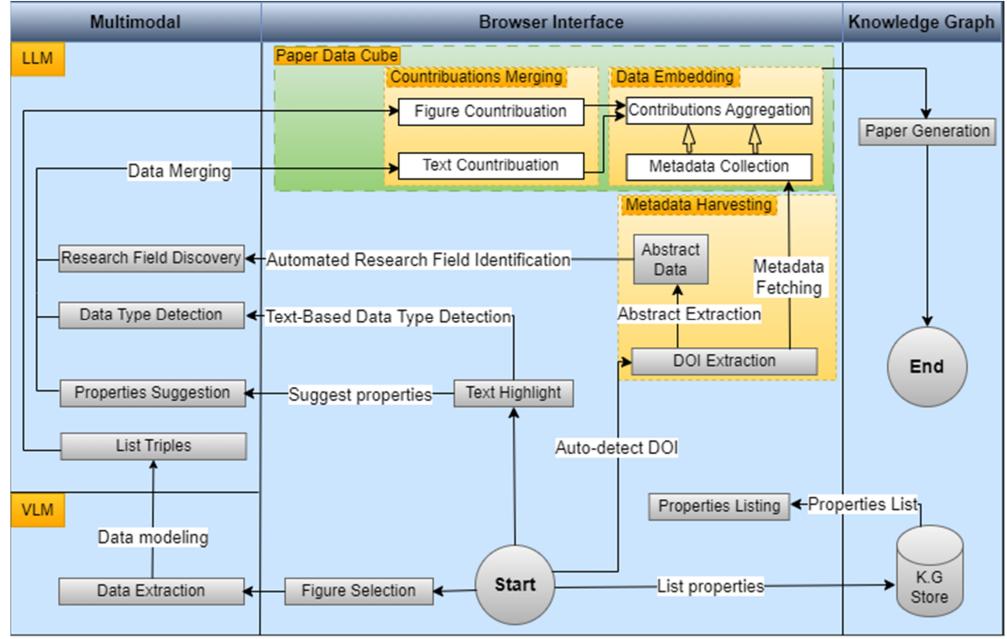




Figure Data Extraction Pipeline

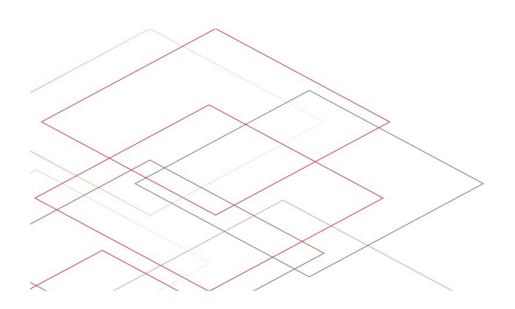




Human-Centric Multi-Modal AI Annotation Pipeline



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	
\times	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

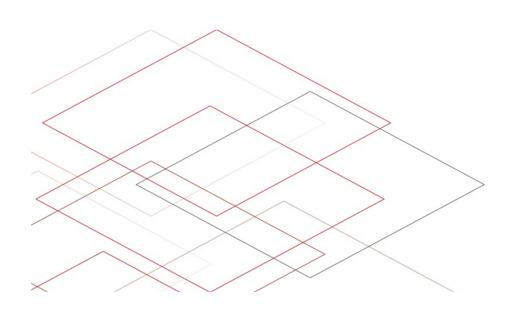
- A. 82% reported faster annotation than traditional interfaces
- B. 90% effectiveness in figure-based triple extraction
- C. 82% trust in Al-generated content, with 73% accuracy perception
- D. High satisfaction (Mean: 4.18 ± 0.75)
- E. Strong Al-assisted performance: metadata extraction (4.27 \pm 0.86), data type detection (4.09 \pm 0.67)

6. Areas for Improvement

- A. Al-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:

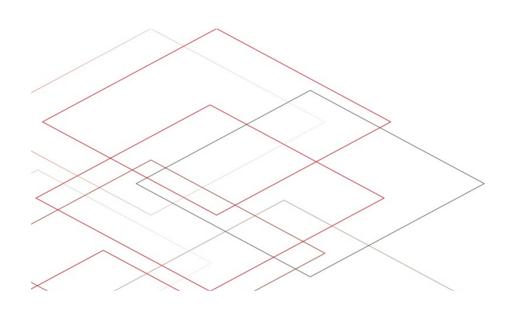
A. Evaluation with 11 participants (82% with ORKG experience)

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





7. Future Work







Future Directions:

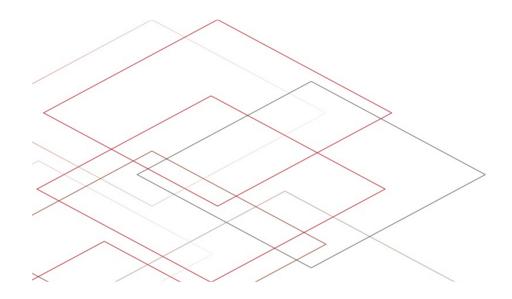
1. Explore advanced techniques in computer vision and NLP (e.g., transformer-based models for figure extraction).

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

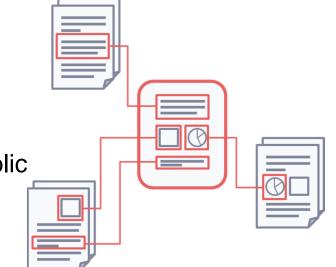
Key Takeaways



Human-Al 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