Does Data Quality Even Matter?

Web of Data Quality Workshop

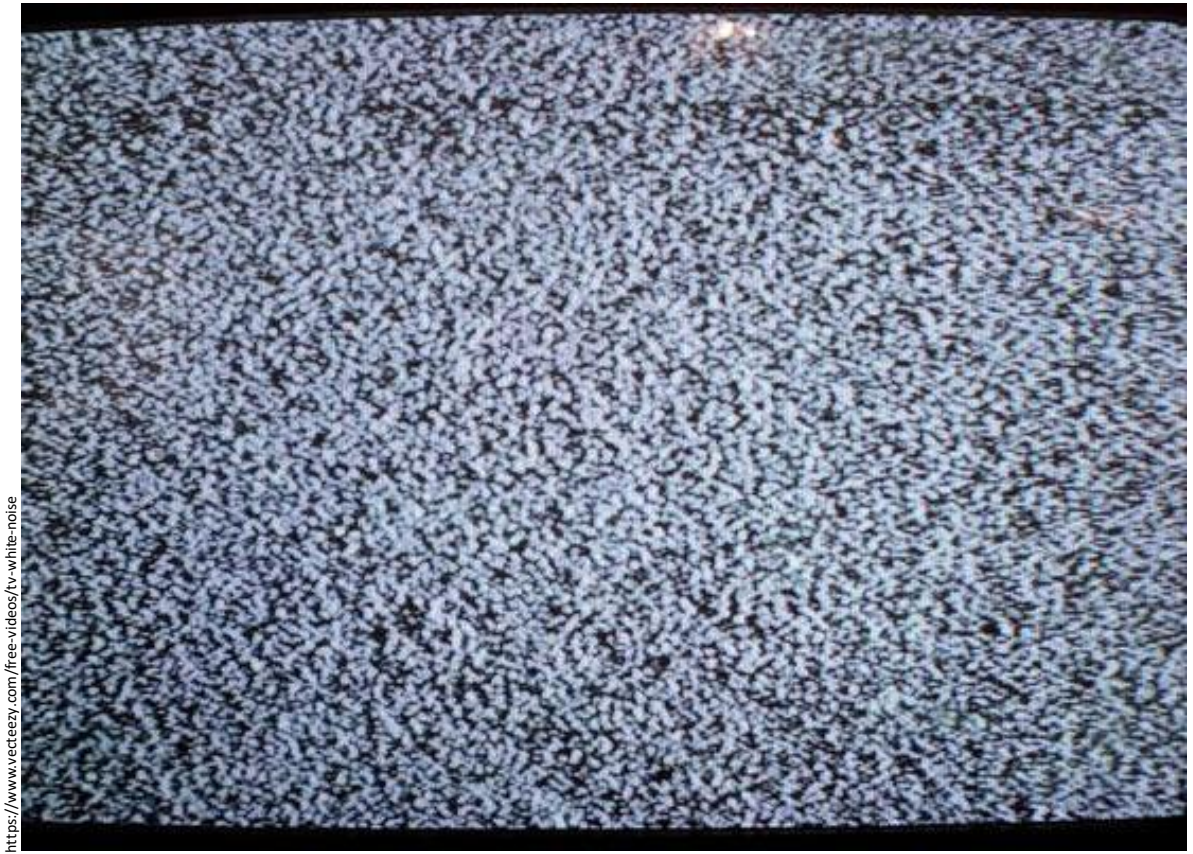
WWW Sydney, 2025

Kerry Taylor

Australian National University

About a decade ago, when deep learning took off in the public eye, everyone was saying that it no longer mattered if you had good data, you only needed a lot of data. And a lot of data there was.

Everyone was putting their data, and their lives, on the Web, and some corporations had the resources to use it all.



https://www.vecteezy.com/free-videos/tv-white-noise

And now we have LLMs.

- It is well known that LLMs can do magic. It is also well known that LLMs *hallucinate*.
- […Trying to swallow my discomfort with such an awful, anthropomorphic word for algorithms making errors…]
- *What happens when LLMs eat their own dog-food?*
- We now know how social media *bubbles* and *echo chambers* can powerfully influence people.
- We also know that in feedback systems the noise can overwhelm the signal.
- *Will this LLM feedback loop kill the Web?*

Models collapse

Shumailov, I., Shumaylov, Z., Zhao, Y. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y

Here we consider what may happen to GPT-{*n*} once LLMs contribute much of the text found online. **We find that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear. We refer to this effect as 'model collapse'** and show that it can occur in LLMs as well as in variational autoencoders (VAEs) and Gaussian mixture models (GMMs). We build theoretical intuition behind the phenomenon and **portray its ubiquity among all learned generative models.** We demonstrate **that it must be taken seriously if we are to sustain the benefits of training from large-scale data scraped from the web.** Indeed, the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of LLM-generated content in data crawled from the Internet.

So what can we do?

Develop Trust infrastructure
Use KGs to help to help the LLMs
Use LLMs to help Ontology Matching
Use data quality measures in KG-RL

ISO-IEC 21617

JPEG Trust

Annotating provenance

Extracting and evaluating Trust Indicators

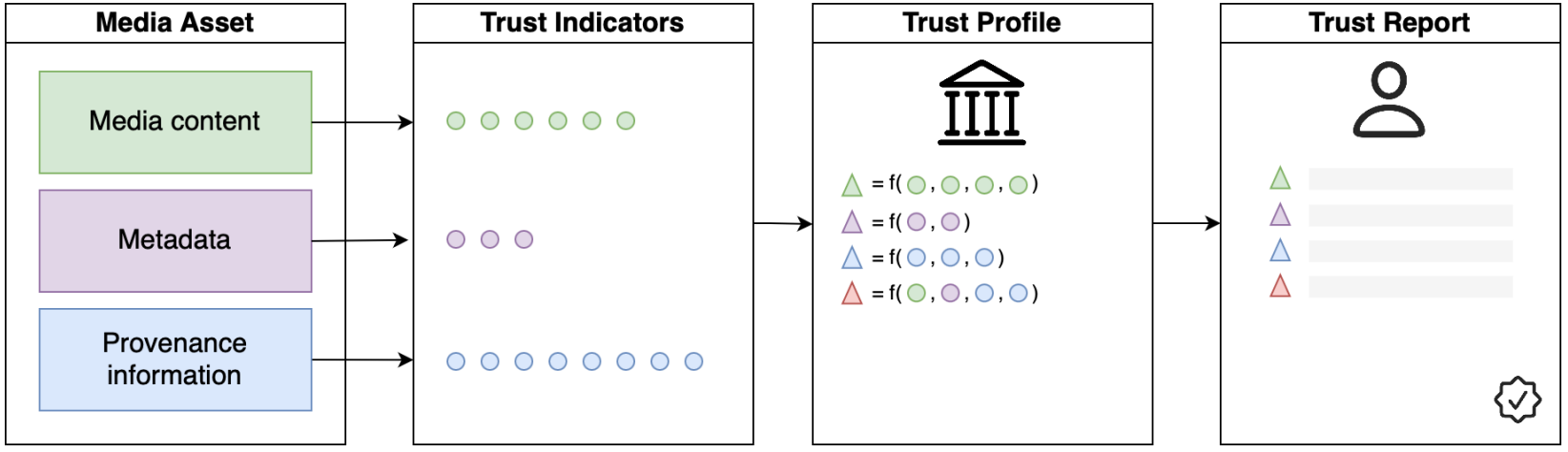
Handling privacy and security concerns

Credit: co-Chair **Sabrina Caldwell** sabrina.caldwell@anu.edu.au

*"The scope of **JPEG Trust** is to provide a framework for establishing trust in media. This framework includes aspects of authenticity, provenance and integrity through secure and reliable annotation of the media assets throughout their life cycle."*

Extracting and evaluating trust indicators

Frederik Temmermans, Sabrina Caldwell, Deepayan Bhowmik, Touradj Ebrahimi,
JPEG Trust: an international standard facilitating the assessment of trustworthiness of digital media assets,
Proceedings Volume 13137, Applications of Digital Image Processing XLVII; 131370B (2024) https://doi.org/10.1117/12.3031171



Status

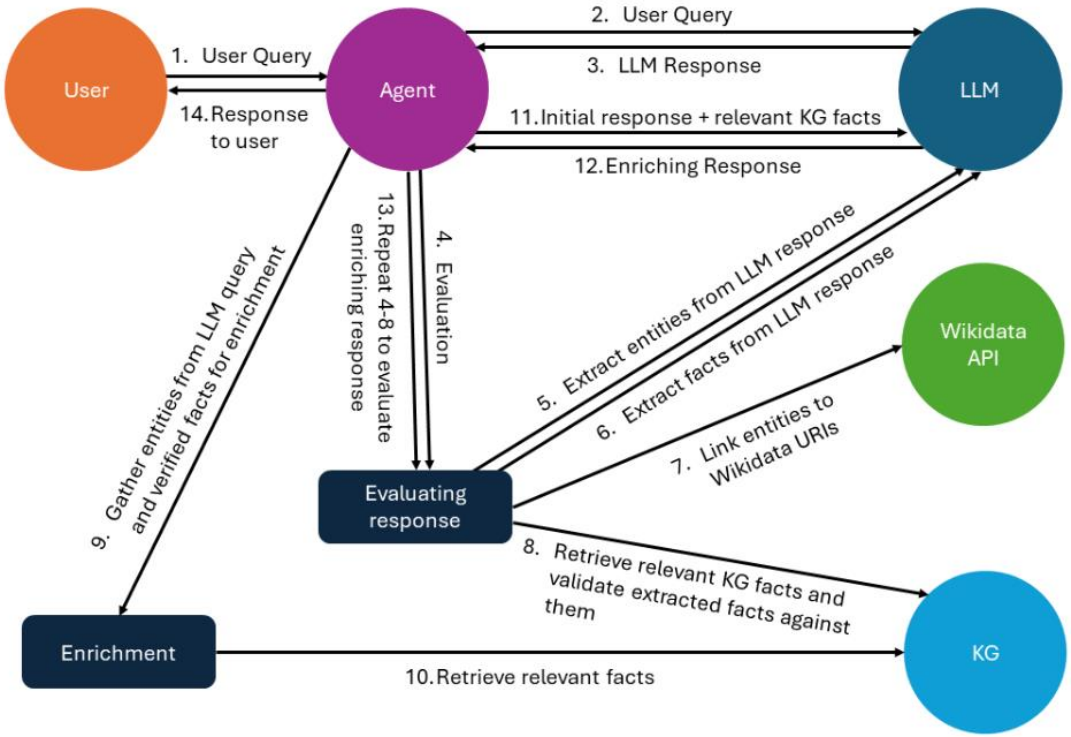
- JPEG Trust Part 1: **Core Foundation** International Standard (IS): July 2024
- Integrated in camera models of Leica, Sony and Nikon.
- JPEG Trust Part 2: **Trust Profiles Catalogue** Working Draft: July 2024

When images, audios, and videos generated by AI are output in files, extended fields must be added to the file metadata for identification. The extended field contains information such as the service provider name, content generation time, and content ID."

foundation models should have information obligations and prepare all necessary technical documentation for potential downstream providers to be able to comply with their obligations under this Regulation. ***Generative foundation models should ensure transparency about the fact the content is generated by an AI system, not by humans****.*

Validating and Enhancing LLM Response with WikiData

Linda Kwan, Pouya Ghiasnezhad Omran, Kerry Taylor [Using Knowledge Graphs and Agentic LLMs for Factuality Text Assessment and Improvement] Posters, Demos, and Industry Tracks at ISWC 2024, CEUR-WS.ORG Vol-3828



Ontology Matching

A very long-term problem to enhance utility and quality of data.

The biggest hindrance to the dream of the Semantic Web: there never was going to be "*One ontology to rule them all*".

Studied for the semantic web since the 2004 at the ISWC conference.

A very similar problem has been studied in Database research since much earlier, as information integration and "schema integration" .

Arguably, the ontology problem is a bit easier, by design.

Then along comes LLMs: Why not?

They can figure out almost anything.

They are eager to please and will give you an answer.

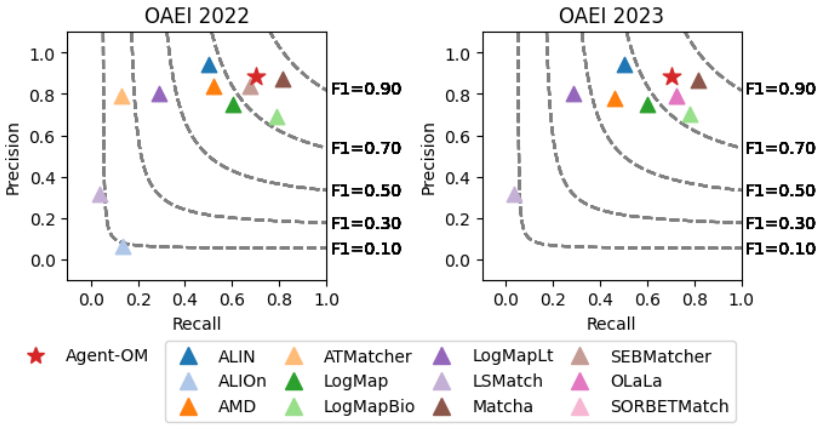
And they write so well!

Zhangcheng Qiang, Weiqing Wang, Kerry Taylor, Agent-OM: Leveraging LLM Agents for Ontology Matching

Proceedings of the VLDB Endowment, Volume 18, Issue 3

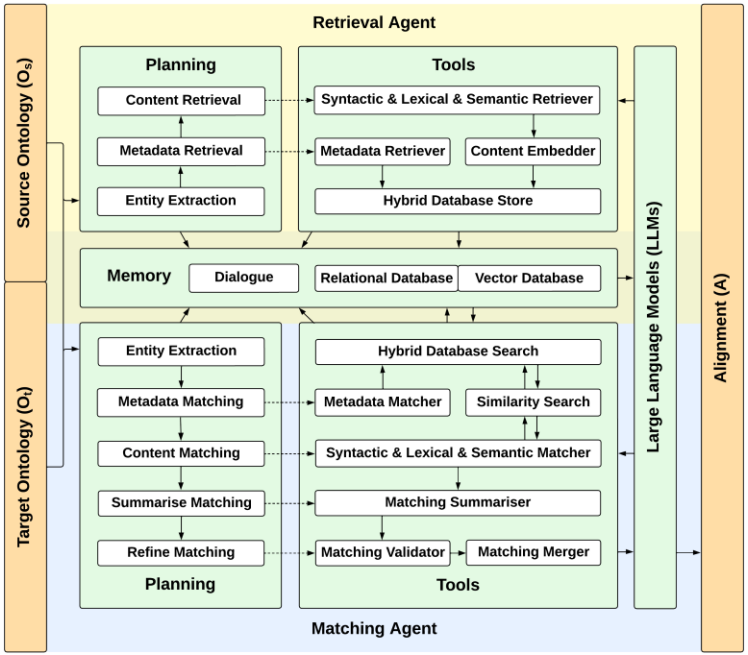
Pages 516 – 529, November 2024

https://doi.org/10.14778/3712221.3712222



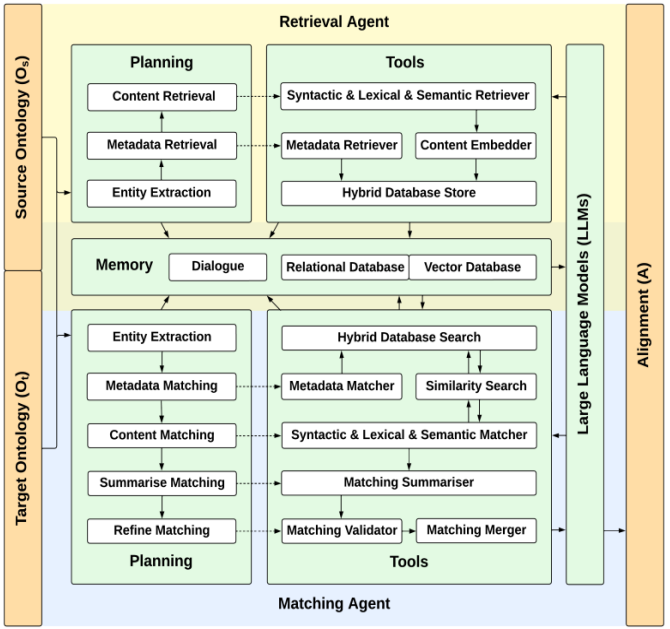
Our work: Agent-OM

- Utilise extended modern features of LLM services
- Pluggable framework for (domain-independent) components including large choice of LLMs
- Use the LLM for planning and coordination, in addition to the usual contextual knowledge
- Employ In-context-Learning and Retrieval Augmented Generation to mitigate hallucinations
- Memory module to communicate and maintain state
- Optimised for cost-effective use of commercial LLMs
- OM performance is leading on complex tasks
- OM performance is equivalent to long-standing conventional performance on simple tasks
- Zero-shot: Needs no "gold standard", "reference", "training data"
- That is, one out-of-the-box Matcher for all OM problems



Contributions

- An ontology matcher that works really well on trivial and especially non-trivial matches.
- Efficient token consumption using vector database similarity search
- Zero-shot, so out-of-the-box ready to go
- A software framework for pluggable LLMs and tool strategies
- An innovative LLM-first design, demonstrating the convenience and robustness of integrated agents
- Stable reliable prompts over varying LLMs (arguably)
- Effective hallucination mitigation (bidirection and self-check)
- Some traditional matchers are better at trivial matching than Agent-OM. Can Agent-OM catch up?



Current Work:

Learning KGs from Rules

Premachandra, A. M., Taylor, K., & Rodríguez-Méndez, S. (2024, November). SPARQL-based relaxed rules for learning over knowledge graphs. In *Proc SEMIIM 2024 at ISWC 2024* CEUR-WS.ORG Vol-3830

Rule-learning over messy heterogeneous agricultural data in a KG.

Want rules for explainability. Dealing with both symbolic and numerical data.

Data quality is highly variable, often relating to the source: carefully designed field trials vs historical field trials vs producer Horn rules.

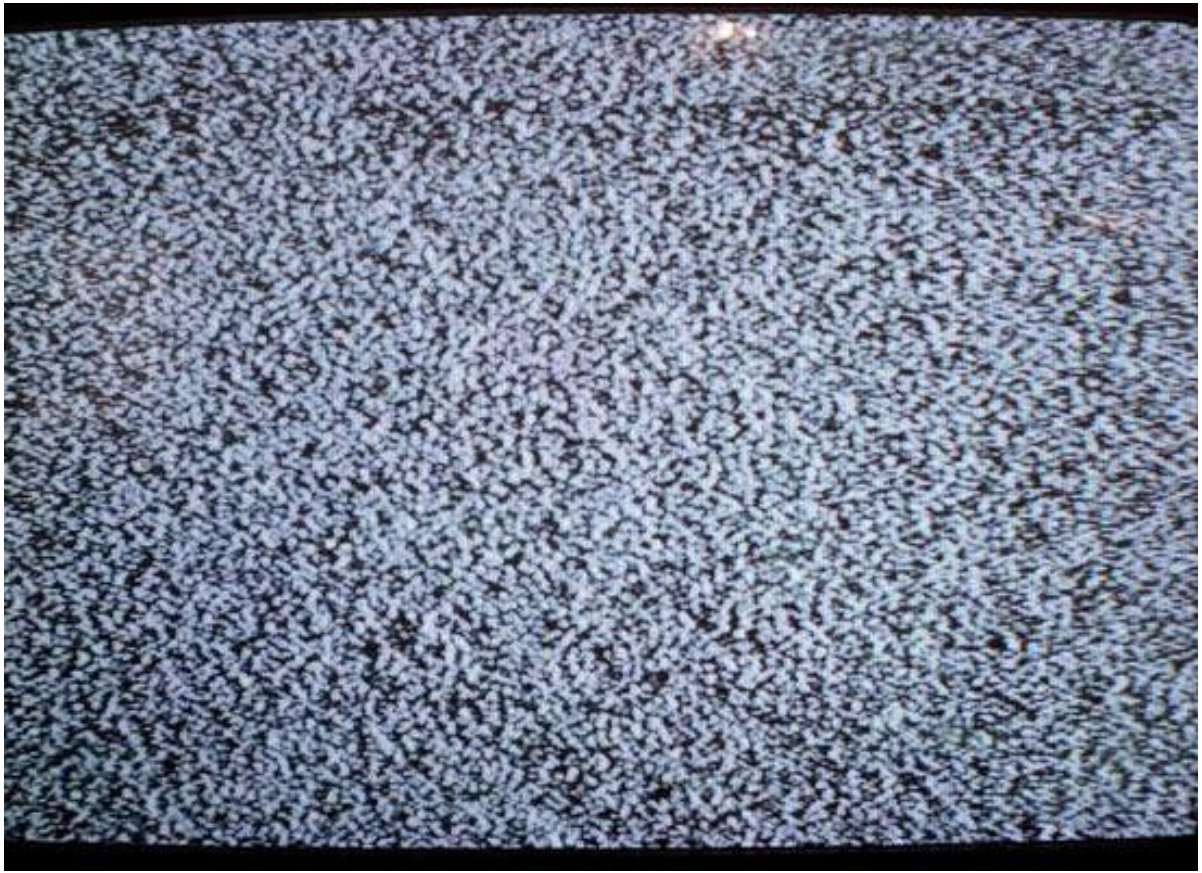
Problem:

Once we have the rules, how do we make recognise the variable quality in the various data sources used to make a (rule-driven) prediction in order to convey confidence in the prediction?

ML needs to understand this better. With rules we have a better chance!

Any ideas?

So does data quality matter any more?



https://www.vecteezy.com/free-videos/tv-white-noise

Thank you

Contact us

Professor Kerry Taylor
School of Computing

E Kerry.Taylor@anu.edu.au
W comp.anu.edu.au

Australian National University

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY)
CRICOS PROVIDER CODE: 00120C