

23rd Plenary Meeting Collaborative Notes

Group(s) name(s) organising the session: [Data Discovery Paradigms IG](#)

Session link: [Applying AI in Data Discovery](#)

Session scheduled date/time/breakout session: Tuesday, 12 November, Breakout 2, 14:30 – 16:00 CST

Venue: Edificio de Ingeniería, Auditorium 2

Session [Slides](#)

<https://www.menti.com> code: 6553 1321

Session summary (for Group co-chairs)

We will use the content in the table below to highlight your work to the RDA community as a report organised by the Technical Advisory Board and to a wider audience through English & Spanish social media mentions.

Please complete ALL fields below by **Friday, 29th November, close of business** to be included in the report & social media activities.

Summarise the session in three sentences:
<i>The session presented three different approaches of using AI to improve metadata quality and user-interface for data discovery, and gathered interest from participants to start a new task force “Applying AI to improve data discoverability”.</i>
Key outcomes/actions/takeaways
1. AI is becoming more relevant to the Data Discovery 2. Users use AI to look for data 3. AI can be used to improve metadata for Data Discovery
Synergies and/or possible collaborations identified with RDA groups and other groups:
FAIR4ML
Highlight text that will be used in social media mentions (please make sure the text is clear and appropriate for public consumption & comprehension)
AI is revolutionising research, including how we curate and discover data. Since RDA Plenary 22, the Data Discovery Paradigms (DDP) IG has been exploring how AI, especially advanced LLMs, can improve data discovery by enhancing metadata, improving interoperability, and refine search interactions. Check out recordings from P23 and P22 for case studies of applying AI to improve data discovery. The DDP IG continues to gather innovative applications and lessons learnt for

the global data community. Join the group and let's unlock the future of data discovery together.

Direct link to Group home page

<https://www.rd-alliance.org/groups/data-discovery-paradigms-ig/activity/>

Get involved in [RDA Community](#)

Check out [P23 programme sessions](#)

This meeting will take place according to the [RDA Code of Conduct](#)

Attendee Check-in

Please complete this table to indicate your attendance (add rows as needed):

Full Name	Affiliation	Location	Email
Chairs: Uwe, Brigitte, Mingfang			
Zachary Trautt	National Institute of Standards and Technology	USA	zachary.trautt@nist. gov
Stéphanie Cheviron	University of Strasbourg	Strasbourg, France	scheviron@unistra.fr
Amir Aryani	Swinburne University of Technology	Australia	aaryani@swin.edu.a u
Anu Gururaj	NIH	USA	anupama.gururaj@n ih.gov
Alex Leykin	Data Belay LLC	USA	databelay@gmail.co m
Christiane Schmidt	GFZ, Helmholtz Geoscience Centre Potsdam	Germany	christiane.schmidt@ gfz-potsdam.de
Gretchen Stahlman	Florida State University	USA	gstahlman@fsu.edu
Piia Markkanen	Oulu School of Architecture, University of Oulu	Finland	piia.markkanen@oul u.fi

Sveinung Gundersen	ELIXIR Norway / University of Oslo	Norway	sveinugu@uio.no
Florina Piroi	TU Wien	AT	florina.piroi@tuwien.ac.at

Collaborative Session Notes *(To be used by participants and chairs during the session)*

Meeting agenda (*)

- Introduction of the group (10 minutes)
- Presentations (30 mins including Q/A):
 - AI/ML to improve data quality (Dr. Amir Aryani, Swinburne University of Technology; Dr. Mingfang Wu, ARDC)
 - AI/ML to improve data discoverability (Dr. Brigitte Mathiak)
- Discussion and feedback on the implications of applying AI in data discovery (40 minutes)
- Wrap up and steps forward (10 minutes)

Discussion topics:

- 1. Please list activities/projects/efforts that really apply AI for data discovery.**

● Please list activities/projects/efforts that really apply AI for data discovery

👤 11

💬 13

Use AI for metadata curation and markup	Traditional search engines for discovery and AI to filter on the quality and usability of the search results
Traditional literature searches	query expansions using ontologies or thesaurus
Augmentation/simulation of data points for sensitive data	Output classification
AlphaFold ?	AI for pooling similar consistent keywords toward a common ontology
Transform questions by users to discovery queries to your search engine	For understanding research terms and looking for literature of research related to ours but not in the core of own expertise
AI mining PIDs	mitigating hallucination
How to make sure the LLMs know about your data?	

2. What are the current gaps you see in the application of AI for data discovery?

● What are the current gaps you see in the application of AI for data discovery?

👤 9

💬 11

Data citation discovery	Metadata usage to leverage AI capabilities
mitigating hallucinations	Searching across repositories
Data to be 'discovered' not described	The challenge of standardization of methodologies for human in the loop validation
Removing thought process for research	People
lack of description verbose enough to be mined	LLMs/GenAI are NOT a search engine, they don't find data, they pretend.
guidance for users to refine searches to favor recent or higher quality versions of data	

3. Which are the possible challenges that you can identify in this context?

Which are the possible challenges that you can identify in this context?

9 13

Who is going to pay for that?	Time to figure it all out.
How best to harness existing metadata for discoverability	Trust
Standards	drift..especially if continuously training
wisdom of the crowd is disappearing, i.e. no conversation on solutions/data is happening anymore.	drift..especially if continuously training
private corporation lockin, not all open (yes some exist)	Off topic, but I worry about the energy consumption. Are we getting enough benefit to justify the environmental cost?
accountability and reproducibility is not possible, even when storing prompts.	integration with other resources
Bias of the GenAI towards the data they've been trained on (mostly English, mostly western), I. E. data availability bias	