Generative AI in Open Research Information processing

Nick Veenstra
Research information specialist, data & LLMOps engineer
University of Groningen / UMCG

About me

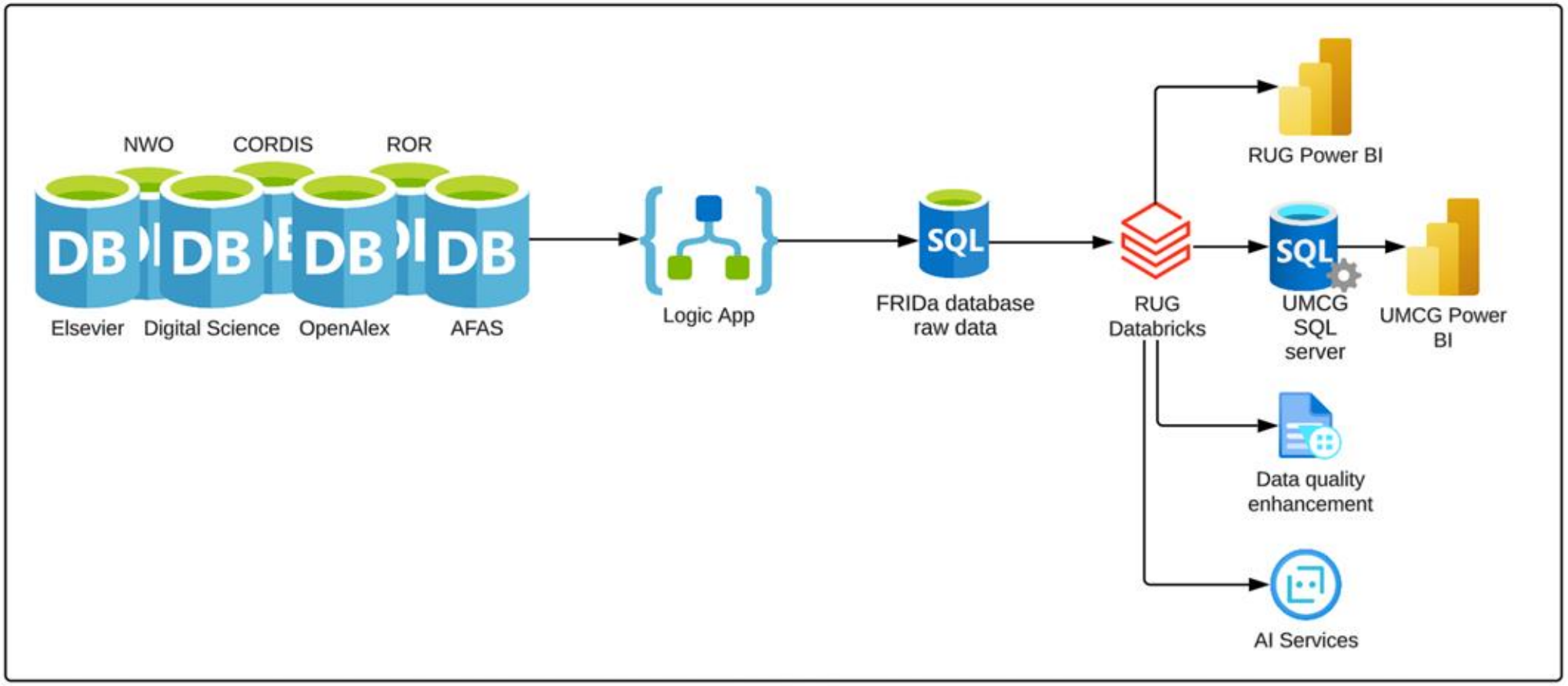
- Librarian
- Library information systems developer (GEAC/Infor - Vubis LIS)
- Research information specialist / Pure manager (TU/e)
- Data / Azure engineer (RUG/UMCG BI project)
- LLMOps engineer (UMCG AI & RUG Pure automation projects)

FRIDa & the rise of Gen AI (2022-2025)

- **F**air **R**esearch **I**nformation **Da**shboard project
  - Joint RUG/UMCG Project to create a local, single source of truth for research information BI
- ‘Local’ data warehouse
  - based on Azure SQL server, Logic apps, Databricks, storage containers
- Collecting data from various sources (18 currently)
- Providing Power BI dashboards for faculty management


What's in FRIDa?
January 2026
5
Altmetric
Scival
ROR
Research Connect (investigating)*
1
Insight/impact
Opportunities
FRIDa
DataWarehouse
4
Pure
Scopus
PubMed
OpenAlex
Web of Science
Espacenet
Result
Funding
CORDIS
NWO
ZonMw
Elsevier grant awards
SNN
Interreg
RVO (investigating)*
2
Projects & data
3
AFAS Profit
Molgenis
DataCite
*: optional data, not required for project

FRIDa architecture overview



AI evolution during FRIDa

During the project:

- ChatGPT arrived
- Databricks' rapid implementation of AI (genie chat and code fixing)
- ChatGPT and Github Copilot integrated in VS Code (dev tool)
- Codex and Cursor as 'AI' first chat IDE
- Dashboards becoming chat interfaces (Databricks One)
- Agentic AI became a thing

This inspired us to start spinoff projects with AI components

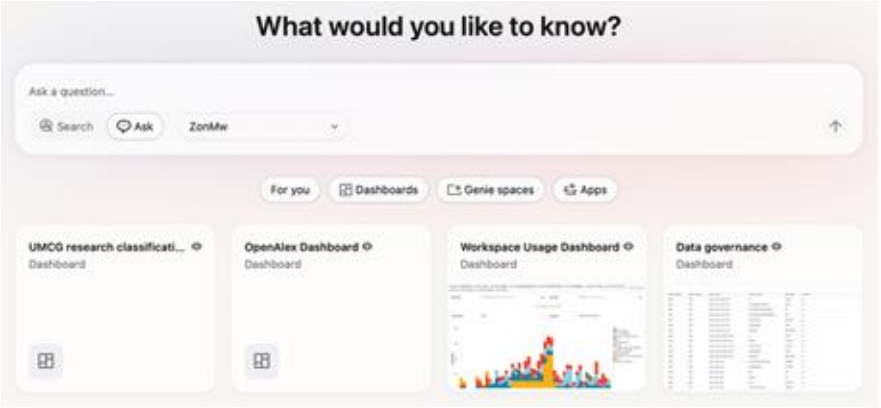
FRIDa spinoff projects

- Automate CRIS (Pure) validation
  - Auto loading of new publications from the data warehouse into Pure (optimal combination of metadata from OpenAlex, (<- 94%!->) Scopus, PubMed and WoS)
  - Mostly process automation, some Agentic AI to maintain data quality

- AI support for UMCG Grant Office
  - Generate newsletters using RAG on funder websites
  - Classify researcher expertise based on output and mission statements
  - Tailor newsletters to expertise/interest
  - AI based chat with data

AI use cases in our projects

- Planning & compliance (ChatGPT)
  - “What’s a good architecture for our data warehouse setup”
  - Get the legal department off our back
  - Digital Autonomy roadmap
- Developing (ChatGPT, Github Copilot, Databricks Assistant)
  - “Give me a SQL script to combine Pure and OpenAlex data”
- Bug hunting (ChatGPT, Github Copilot, Databricks Assistant)
  - “Hey ChatGPT, why doesn’t the SQL script that Databricks’ assistant gave me work?”
- Data quality fixing (ChatGPT, Github Copilot)
  - Add more DOIs, ORCiDs, clean up external organizations
- Chat with data (Databricks Genie & One)
  - AI instead of Power BI to explore the data
- Classifications (Local LLM; gpt-oss-120b on DGX Spark)
  - Build and apply a medical taxonomy for UMCG research



Autonomy backup plan (AI assisted)

Current solution	Backup solution
Azure SQL server	Local PostgreSQL (better for AI anyway)
Azure Logic apps	Local Jupyter notebooks
Azure Databricks orchestration	Databricks on AWS or Google, (Lidl?) or local Apache Airflow
Databricks notebooks	Local Jupyter notebooks
Azure Devops version control	Local GitLab
Mistral / OpenAI AI	Local LLM’s
Power BI	Apache Superset

Improving data quality with AI

Loading and fixing data need to work in tandem

- Use of AI to plan, code and bulk fix (e.g. external organizations)
- Process automation to load new data (e.g. publications,ROR organizations in the CRIS)
- Agentic AI to maintain data quality and cleanup (due to lacking control in the interface options)

ai_ror_id	ai_confidence	ai_notes
	low	The name 'Nederlands Fries Nedersaksisch' does not point to a specific organization. It seems to reference the Dutch Frisian and Low Saxon languages rather than a distinct institution. Without additional context, it's challenging to associate this with an organization.
https://ror.org/03p7fer83	high	The entry mentions multiple associated institutions within the same research unit context. The ROR ID corresponds to the LPC, which is a recognized unit under CNRS/IN2P3 and Blaise Pascal University.
https://ror.org/05dxps055	high	The name 'Sez INFN Ferrara' matches well with the INFN section located in Ferrara, Italy.
https://ror.org/022g41249	medium	The provided entry 'Grieks, Latijn, Mediaevistiek en Nieuwgrieks' seems closely related to departments or units found within Leiden University in the Netherlands, known for humanities and classical studies. However, due to the lack of a specific city and multiple organizations in the Netherlands potentially matching this description, confidence is medium.
https://ror.org/05xpvk416	medium	The Institute for High Energy Physics (IHEP) is a well-known institution, typically associated with Protvino, Russia, rather than the United Kingdom. The discrepancy in the country may affect the confidence level.
https://ror.org/041b64y08	high	The Alikhanov Institute for Theoretical and Experimental Physics is a well-known entity in Russia, often referred to by its acronym ITEP, and matches the provided name closely.
https://ror.org/00x1avg12	high	The entry mentions multiple organizations, but the Laboratory of Physics of Clermont (LPC), which is a unit of both the CNRS and IN2P3, aligns well with the provided details, including the location in Clermont-Ferrand.
https://ror.org/0563far46	high	The organization name is specified with the department and hospital name, and there is a direct match in the ROR database with a corresponding ROR ID for Tuen Mun Hospital in Hong Kong.
https://ror.org/03rh3cw74	high	The organization combines the University of Aix-Marseille, CNRS, and specifically refers to CPPM and IN2P3, which aligns with CPPM being part of CNRS-IN2P3.
https://ror.org/013meh722	high	The name 'Centro Brasileiro de Pesquisas Físicas' matches closely with the entry provided. The organization is based in Brazil, aligning with the 'Brasileiro' in the name.

Generating with AI

AI offers the opportunity to create customized classifications, abstracts for CRIS and BI

- can vary slightly across runs (also goes for humans)
- may shift behavior across model updates
- may re-interpret low quality data differently over time

January 29, 2026 Product

Retiring GPT-4o, GPT-4.1, GPT-4.1 mini, and OpenAI o4-mini in ChatGPT

LLMOps = stability

Consistent data quality, prompting, model testing, schema's, taxonomies needed

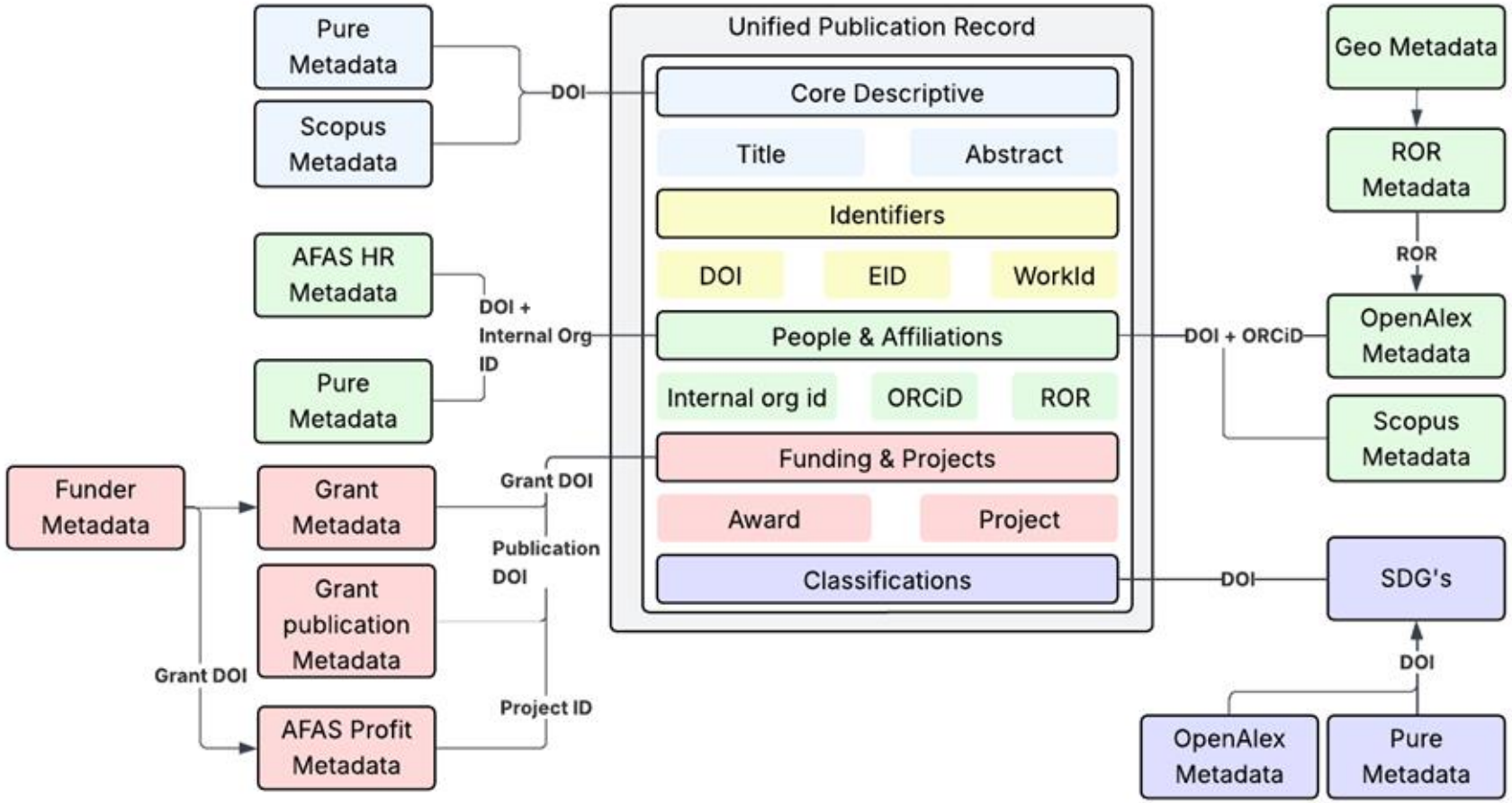
Sample

		year	classification
1	ers is rising internation...	2024	{"health_categories": [{"name": "musculoskeletal", "confidence": 0.9}, {"name": "generic-health-relevance", "confidence": 0.7}, {"name": "neurological", "confidence": 0.5}], "research_activity_subcodes": [{"id": "2.4", "name": "Surveillance and distribution", "confidence": 0.85}, {"id": "1.4", "name": "Methodologies and measurements", "confidence": 0.75}, {"id": "8.4", "name": "Research design and methodologies (health services)", "confidence": 0.6}]}

Operational AI Stack for Research Libraries

Component	Goal	Maintainers
Components and orchestration	Reusable models, agents, MCP servers, microservices, pipelines	LLM Ops, integration specialists
Systematic benchmarking	Uniform monitoring of different models (performance, quality)	Librarians
Capabilities registry	Registry of AI capabilities (tooling, templates, features) with cost profiles, benchmark scores and usage guidelines	LLM Ops / librarians
Standardization	Common system prompts, data models and definitions, guardrails	Librarians

The need for unified data instead of vendor silos



Conclusions

- AI enables rapid prototyping and development, flexibility in data autonomy and sovereignty

- New use cases: loading, fixing, classifying and chatting with the data
  - AI will probably (in part) replace current analytics interfaces
  - Improve data quality with (agentic) AI (essential with diminishing library staff)

- Librarians need to focus more on prompt engineering, model selection and testing to allow data curation at current scale

- Focus on infrastructure (national):
  - Quality, curated data not bound/structured by vendors
  - Stable and autonomous AI + expertise sharing

Thanks! Questions?