HELMHOLTZ Metadata Collaboration

AST | E&E | ENERGY | HEALTH | INFORMATION | MATTER | DATA COMMONS | HMC OFFICE | HMC PROJECTS

Connecting information across repositories – a keyword-based approach

Stanislav Malinovschii (GEOMAR) // Emanuel Söding (GEOMAR) // Andrea Pörsch (GFZ) // Dorthee Kottmeier (AWI) // Yousef Razeghi (UFZ) // Sören Lorenz (GEOMAR)

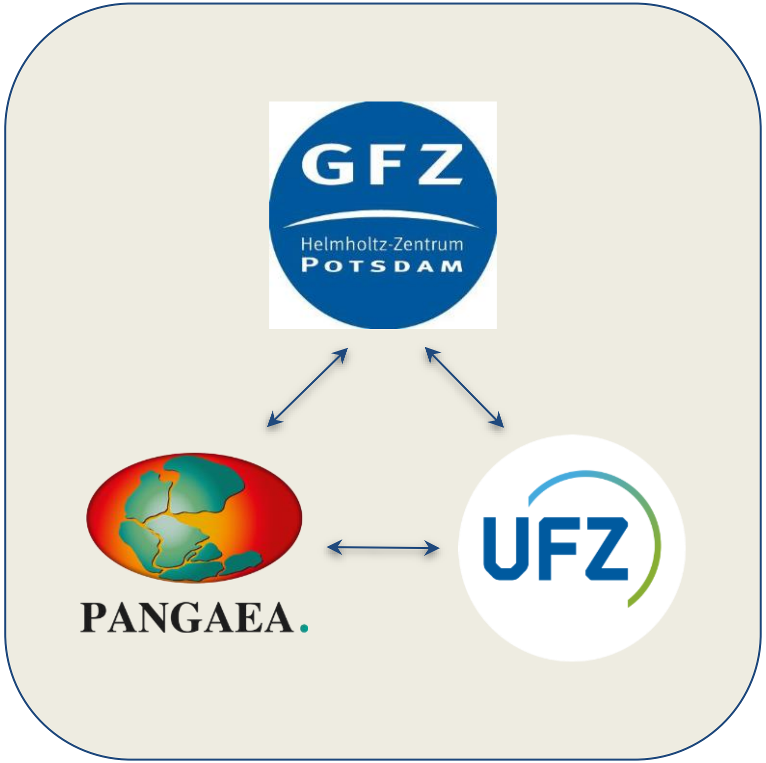
www.helmholtz-metadaten.de

Introduction

Knowledge graphs help integrate and structure information from various sources and entities, enabling advanced search and filtering techniques across large datasets to reveal hidden connections and dependencies. However, achieving this integration **requires uniform and harmonized datasets**.

Currently, most knowledge graphs in scientific research are **based on bibliographic data**, which limits their utility for scientific purposes due to a lack of substantive content.
To enhance scientific relevance in Earth and Environmental research, we focus on identifying key parameters within data metadata to build a more comprehensive knowledge graph.

In this slide, we present our approach to gathering and analyzing metadata from several Helmholtz repositories. We discuss the opportunities and challenges in creating knowledge graphs and provide statistics on the collected data along with recommendations for improving data quality.



Strategy and Approach

Main Objective:

Our primary goal is to create disciplinary knowledge graphs using Neo4j to connect scientifically relevant information, such as methods, instruments, measured parameters, and more.

This approach enables us to develop a structured network of scientific information that enhances the accessibility and utility of domain-specific knowledge.

Repository Selection

Initially we select some repositories: e.g. Pangaea, GFZ, UFZ and Hereon institutional repositories

Metadata Harvesting

Using the OAI-PMH and CSW protocols, we harvest metadata in the different XML schemas from these repositories.

Keyword Parsing: Extract Scientific Terms.

We then parse all keywords from the metadata to identify critical scientific terms and descriptors.

Keyword Grouping by Related Terms.

Next, we organize these keywords into meaningful groups to establish connections between similar or related terms.

Knowledge Graph Creation

Finally, we use Neo4j to build the knowledge graph. That way we link these grouped keywords to form a cohesive structure to connect disciplinary data.

Challenges

We tested our approach by harvesting keywords from the GFZ and Pangaea databases:

- Pangaea: 206,605 unique keywords
- GFZ Data Services: 7,154 unique keywords

This presented the following challenges:

- Keywords are based on different vocabularies, if any.
- Some keywords lack clear meaning without additional context.
- Aligning and mapping various semantic meanings is complex and time-consuming.
- Categorizing keywords is challenging and requires considerable time.

Challenges

Example from Conductivity - Temperature - Depth Probe (CTD):

ctd	ctd from ice float
ctd (sbe19)	ctd casts
ctd (sbe19) and niskin bottles (8-l or 12-l) triggered with messengers	ctd, ictd, sn 1360
ctd (sbe9s)	ct, rbr, rbrduo c.t
ctd 60 (sea & sun technology gmbh, germany)	ct-probe aqua **troll** 100
ctd 60m multiparameter probe (sea & sun technology gmbh, germany)	ct-scan
ctd probe	ct-scans

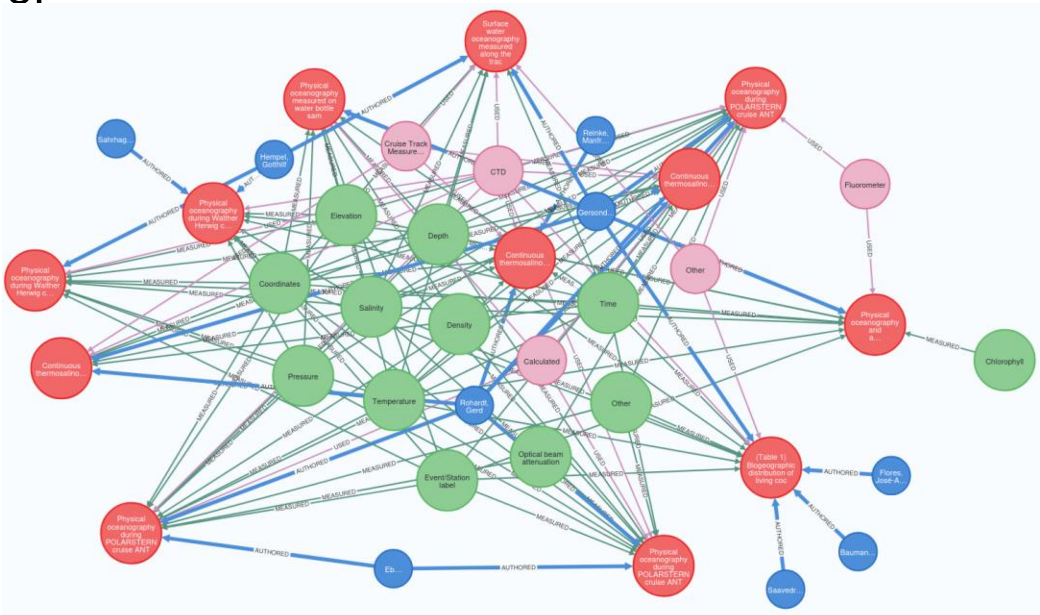
06zg20100207
06zg20101023
07al692_2
08bd0394_1
096u20160630
09ar0103
09fsh02

This volume of keywords with overlapping or synonymous meanings presents a serious challenge to effectively organize and harmonize data within the knowledge graph, turning it simply into a large dumping ground of disparate parameters and isolated nodes.

Current Standing

At the moment we have defined some rather general but capacious categories, which we will develop and expand in the future:

- Projects
- Platforms
- Locations
- Disciplines
- Methods
- measured Parameters
- Instruments



We also built a small knowledge graph based on a CTD data subset from PANGAEA, which showed that our approach “inside” the repository gives good results: with ~23000 nodes we have ~245000 links.

Next Steps

1. Test methods to sort keywords into meaningful categories comparing LLMs to manual methods
2. Mapping keywords into vocabularies or taxonomies
3. Annotate a limited number of datasets from different repositories with meaningful disciplinary keywords.
4. Then build a **small knowledge graph** focused on a category that connects scientific repositories using harmonized lists of keywords to test our procedure
5. Expand the list of keywords by parsing publication abstract from metadata schemas.
6. Leverage more repositories.
7. Build a large knowledge graph that connects scientific repositories using harmonized lists of extended keywords.


THANK YOU FOR
YOUR ATTENTION!