AI and Metadata

What really changes?

Andrea Scharnhorst | Senior Researcher
Jetze Touber | Data Station Manager Humanities
Zehao Lu | Research Engineer

June 11th, 2026

DANS Open Day
Open data, open science

DANS

Welcome and introductions

DANS Strategy 2026

Responsible use of AI in data services (enrichment and findability of metadata and datasets)

Role of data experts and curators for **safeguarding** integrity, quality, and validity of (meta)data.

Generative AI questions about intellectual property, transparency and traceability of knowledge, and possible conflicts with **principles** of open science

Jetze Touber - DANS - Data Station Manager - SSHOC-NL

Zehao Lu - WUR - Beijing University/Utrecht University - LTER-LIFE, AI researcher

Welcome and introductions

Demonstrations (15')

SSHOC.NL NLP and LLM to autosuggest structured keywords

LTER-LIFE Metadata Agents

Discussion (15')

Your experiences and expectations

Quality control and data provenance:
human-in-the-loop, or not?

Legal issues:
Exposing unpublished data to AI-services

Infrastructural needs:
Which AI-stack is available and desirable?

Closing Statement

Results Agentic AI workflow (LTER-LIFE)

Jetze Touber

DANS

AI and Metadata

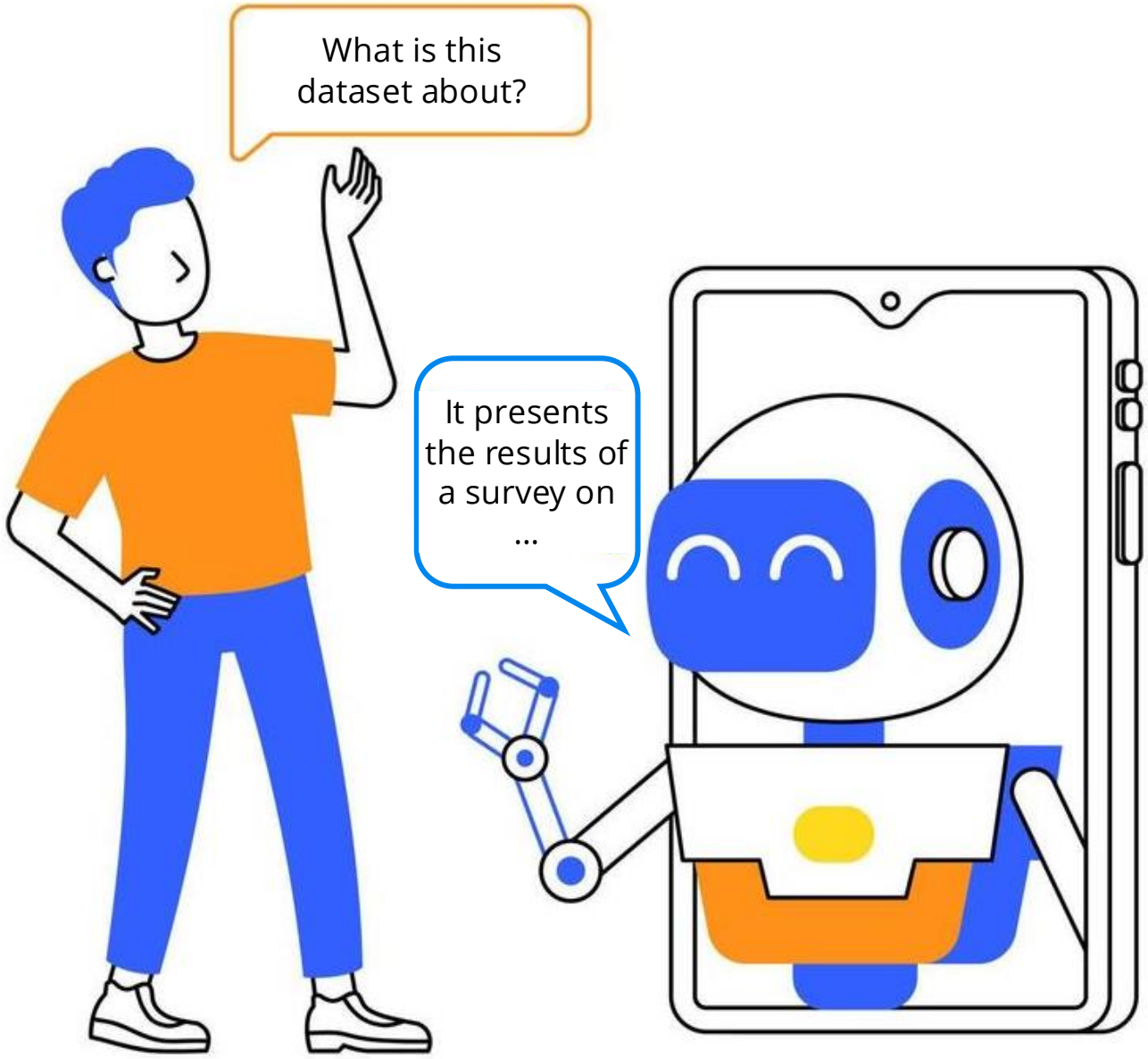
Dataset is a black box

Metadata describe what's in the box

AI and Metadata

Dataset is a black box

Metadata describe what's in the box



Potential of AI-methods

1. Improve metadata
2. Extract metadata from data
3. Interact with data directly

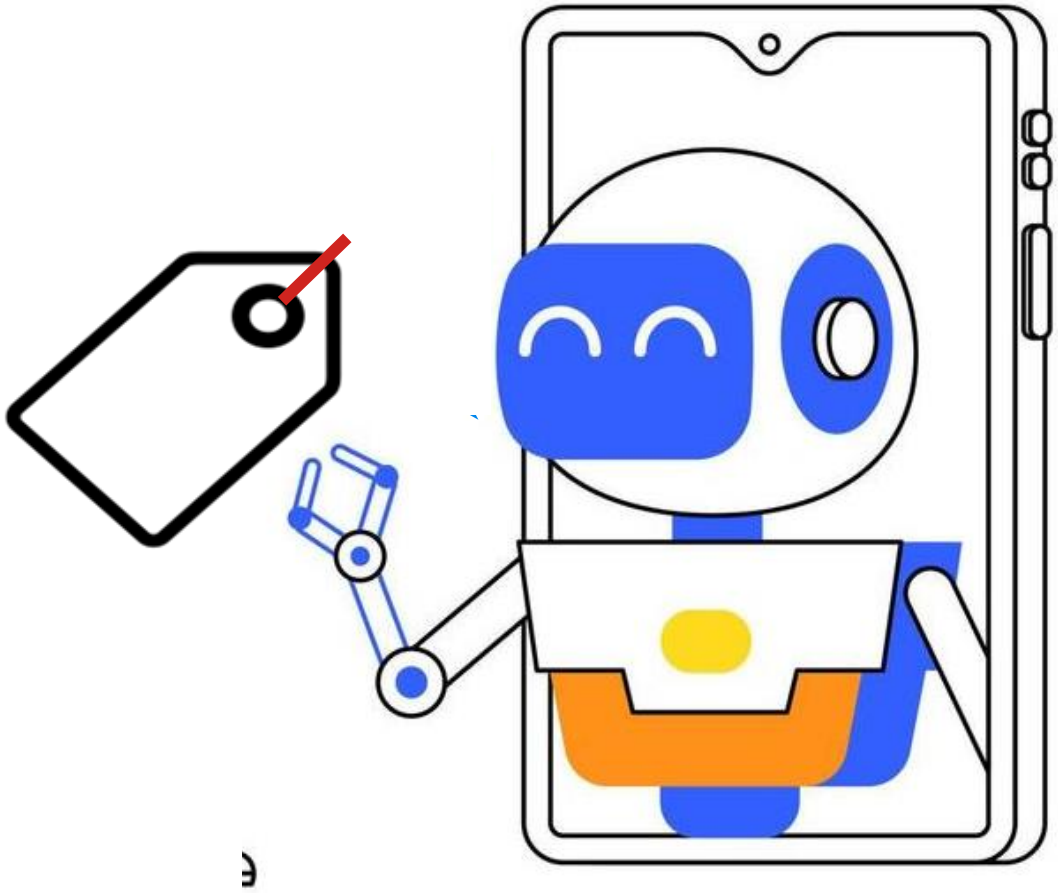
AI and Metadata

Dataset is a black box

Metadata describe what's in the box

Potential of AI-methods

1. Improve metadata > **e.g. keyword selection**
2. Extract metadata from data
3. Interact with data directly

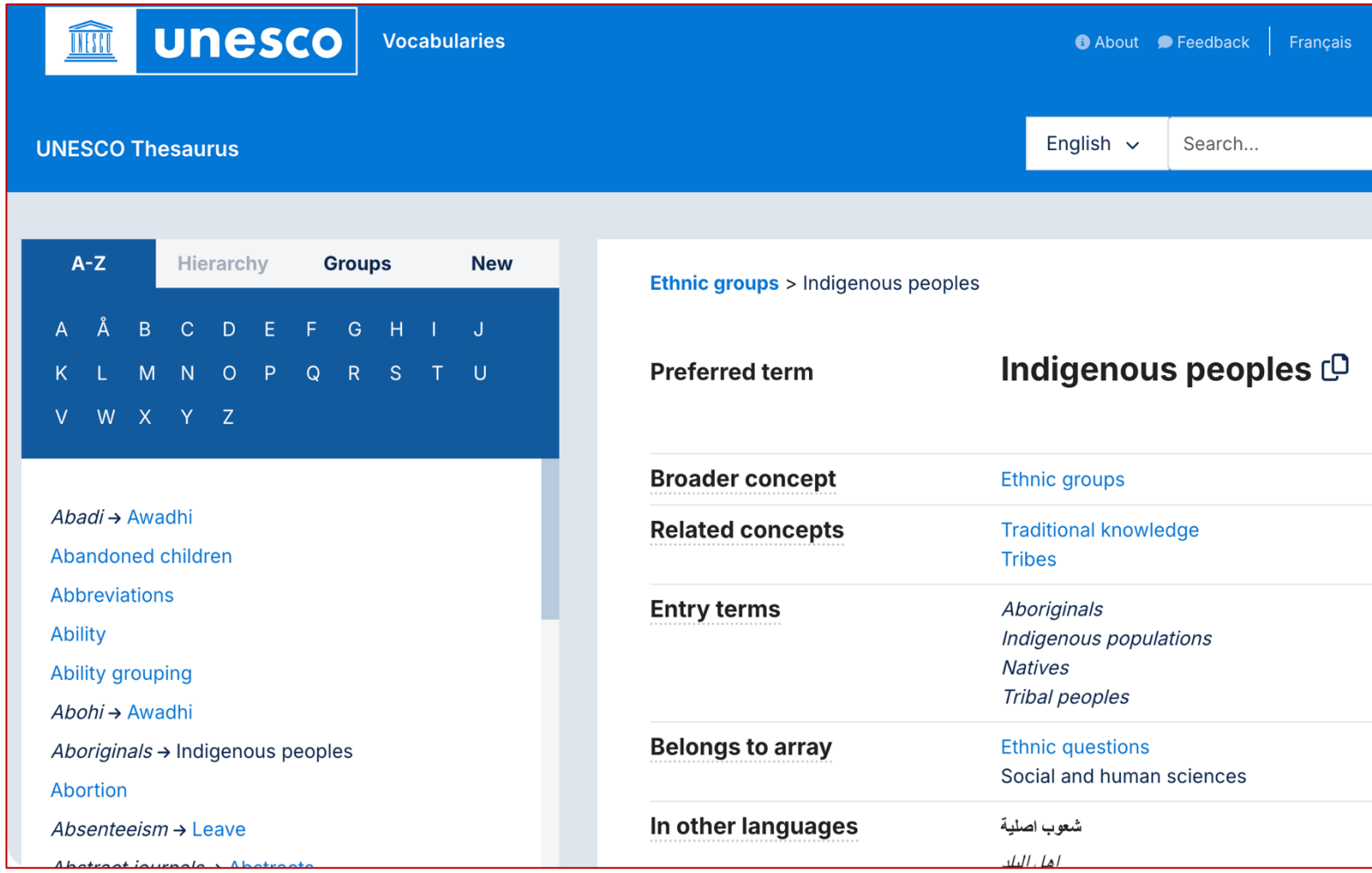


Automated Subject Indexing (ASI)

- Keywords to describe content (**index**)

Automated Subject Indexing (ASI)

- Keywords to describe content (**index**)
- From **controlled vocabulary**

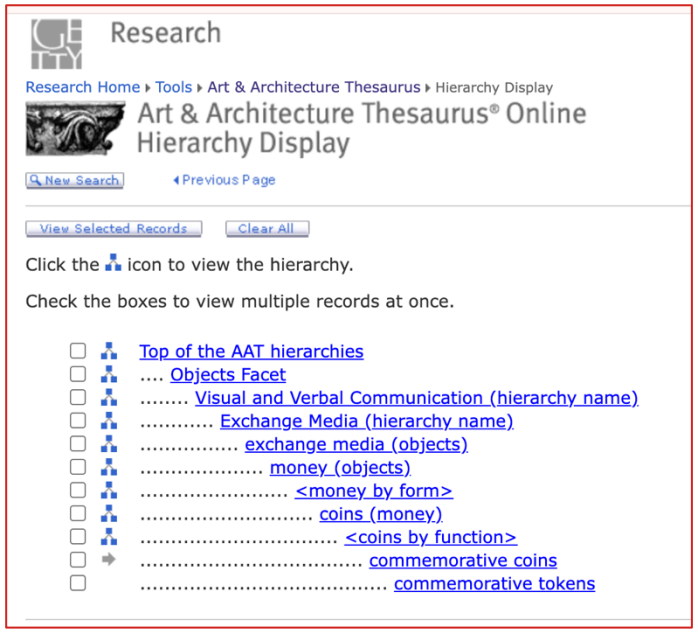


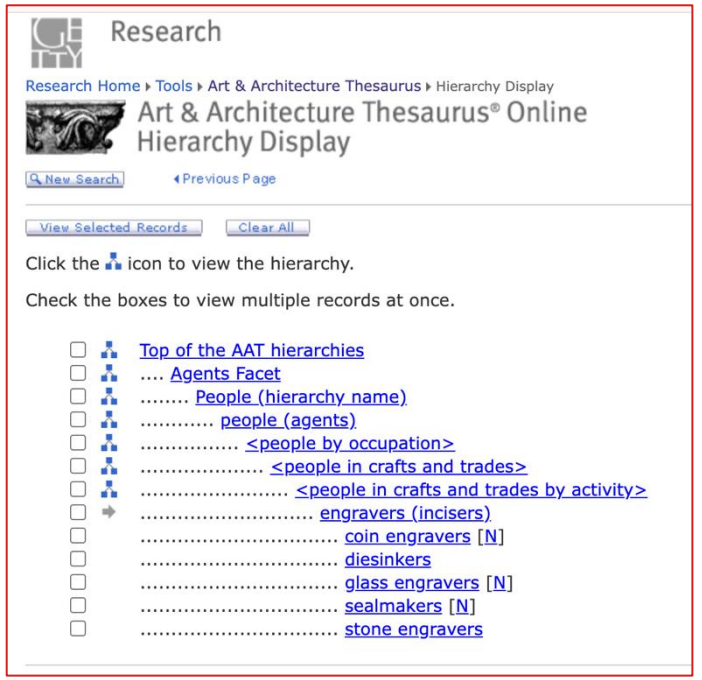
Automated Subject Indexing (ASI)

- Keywords to describe content (**index**)
- From **controlled vocabulary**

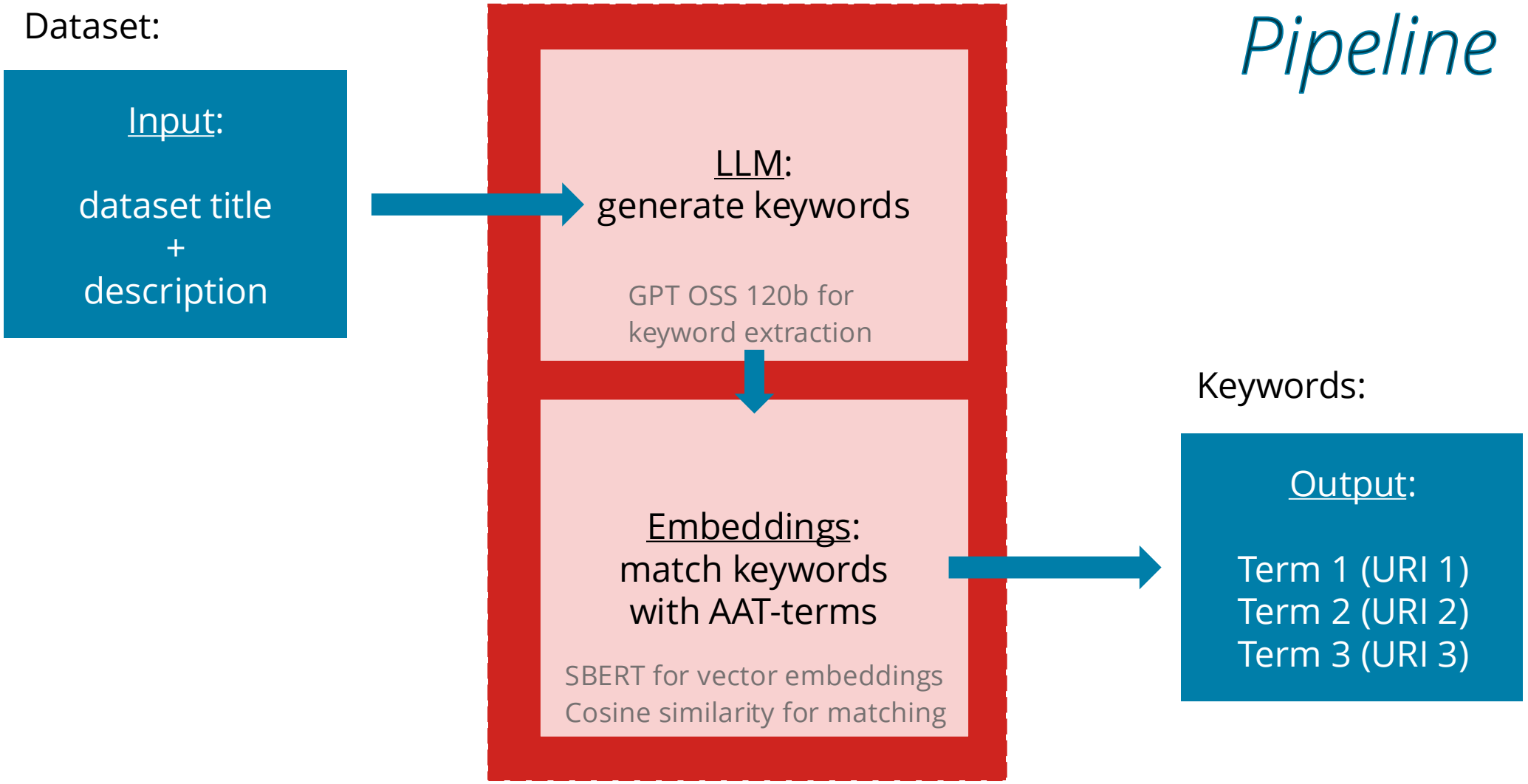
Experiment:

Autosuggest terms from Getty's Art & Architecture Thesaurus (**AAT**)





Automated Subject Indexing (ASI)



Pipeline

Automated Subject Indexing (ASI)

Dataset:

TH. Vermaut, 2018, "Gebuurtenkaarten/Huisnummerkaart Leiden, omstreeks 1853", https://doi.org/10.17026/DANS-XGS-BTP3, DANS Data Station Social Sciences and Humanities, V2

Title

Gebuurtenkaarten/Huisnummerkaart Leiden, omstreeks 1853

Description

Zie hiervoor het bijgevoegde meta-data file.

Generated keywords:
"historical map"
"street map".
"19th century"

Matched AAT-terms:
"historical maps"
"road maps"
"nineteenth century (dates CE)".

Example

Keywords:

"historical maps" (+ URI)
"road maps" (+ URI)
"nineteenth century (dates CE)" (+ URI)

Reflections

Benefits

- **Nudge** depositor
  to use a controlled vocabulary
- **Assist** depositor
  in selecting from a huge number of terms

Concerns

- Relevance is *subjective*
- Requirements for *input*?
- Expose *unpublished* materials

> Depositor must be in control
> Transparency is key

Zehou Lu

WUR

Metadata from Dataset

nnual_budburst_per_tree.csv

rganismID	bud_burst_date	bud_burst_DOY	verbatimLocality	scientificName
1	NA	NA	Hoge Veluwe	Quercus robur L.
2	1988-05-01	122.5	Hoge Veluwe	Quercus robur L.
3	1988-05-03	123.666666666667	Hoge Veluwe	Quercus robur L.
4	1988-05-01	122.5	Hoge Veluwe	Quercus robur L.
5	1988-05-01	121.8	Hoge Veluwe	Quercus robur L.
6	1988-04-28	119	Hoge Veluwe	Quercus robur L.
7	NA	NA	Hoge Veluwe	Quercus robur L.
8	1988-04-28	119	Hoge Veluwe	Quercus robur L.
9	1988-05-03	123.666666666667	Hoge Veluwe	Quercus robur L.
0	NA	NA	Hoge Veluwe	Quercus robur L.
1	1988-04-28	119	Hoge Veluwe	Quercus robur L.
2	1988-05-01	122.5	Hoge Veluwe	Quercus robur L.
3	NA	NA	Hoge Veluwe	Quercus robur L.

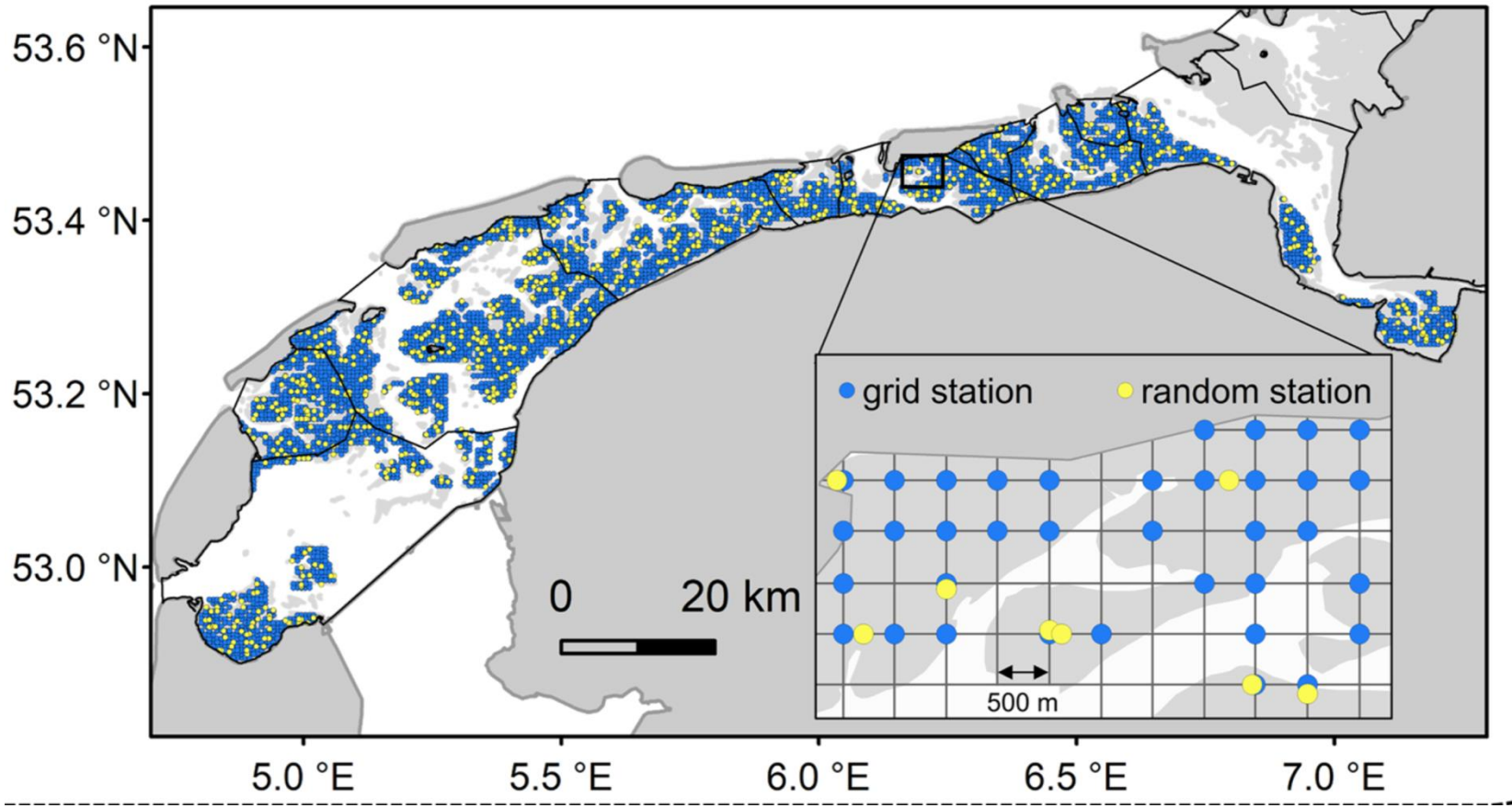
Metadata Agent

1. Dataset with missing metadata @SURF RDM
2. (Before) Researchers are required to complete metadata forms when uploading datasets.
3. Can we use GenAI (LLM) to identify those information?
   - What is the dataset about
   - What is its spatial and temporal coverage?
   - What is the resolution?

Metadata Agent @Wageningen University @SURF

Metadata Agent

1. LLM Harvestor (LTER LIFE)
2. Metadata with GenAI
3. Agents
   - LLM is not reliable
   - Proper system design allows us to build reliable system with unreliable components
   - Tools, Planning, Memory
   - *LLM can talk, Agents can do.*



SIBES: Long-term and largescale monitoring of intertidal macrozoobenthos and sediment in the Dutch Wadden Sea @NIOZ

Thank you for your attention

DANS Open Day
Open data, open science

DANS

Anna van Saksenlaan 51 | 2593 HW The Hague | The Netherlands | +31 (0)88 003 46 66
National centre of expertise and repository for research data | An institute of the KNAW and NWO

DataLink LinkedIn BlueSky Mastodon www.dans.knaw.nl