

Linking Agrosystem Data with Socio-economic Information

Use Case Pilot (12 months, March 01, 2025 till February 28, 2026)

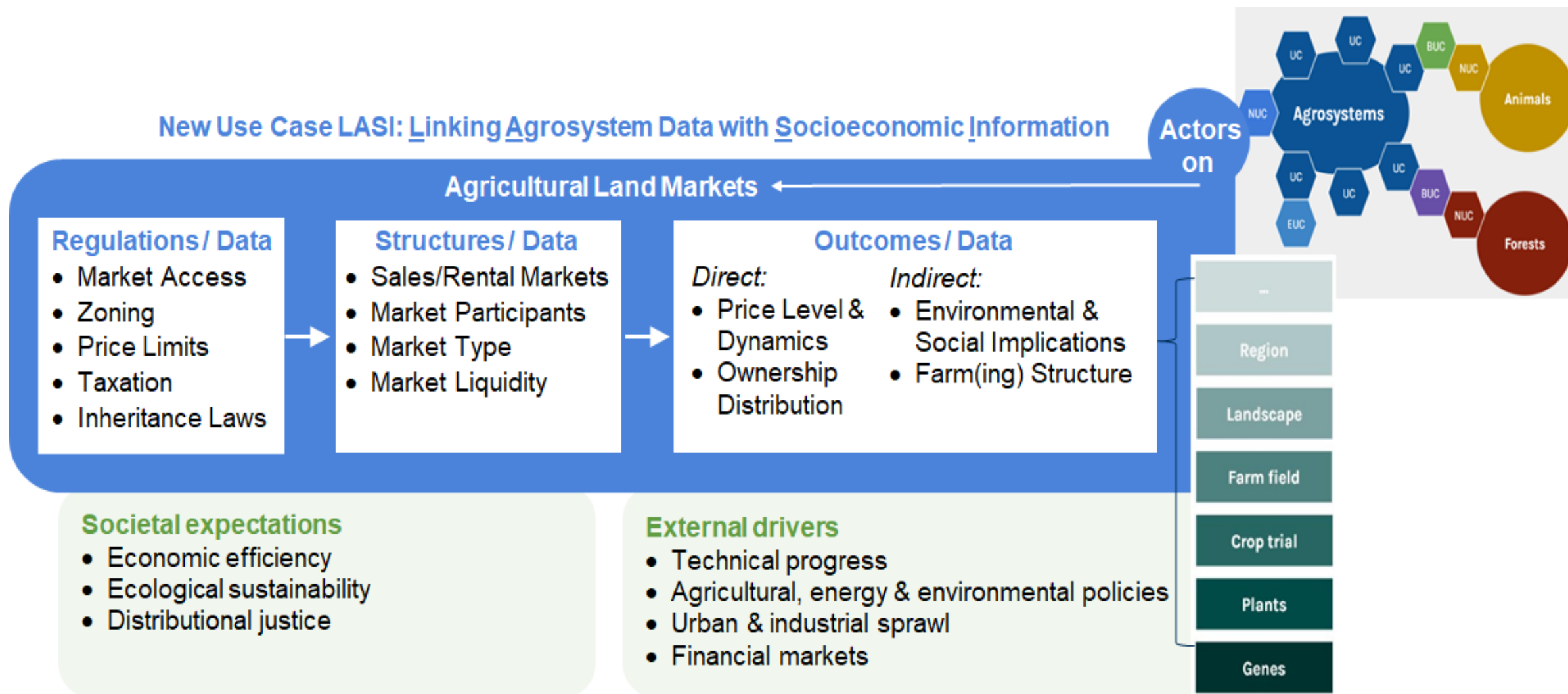
Farmland Data – Structures, Challenges and FAIR Principles

Lorenz Schmidt, Dr. Günther Filler, Prof. Dr. Timo Kautz, Prof. Dr. Martin Odening

Humboldt-Universität

Associated Partners

Alfons Balmann (IAMO Halle), Oliver Mußhoff (University of Göttingen)



Source: Forland <https://www.forland.hu-berlin.de/en/institut-en/departments/daoe/forland>; Fairagro; own compilation

Key stakeholders and their interests in agricultural land market data

- **Farmers**
 - Ask for access to fair and competitive land prices
- **Investors**
 - Assess investment opportunities, mitigate risks associated with agricultural land investments
- **Government Agencies**
 - Demand for evidence-based data for policymaking, effective land market Regulation
- **Researchers and Academics**
 - Access to structured datasets to (e.g.) provide policy advice, carry out an economic evaluation of farms and land use systems
- **Environmental Organizations**
 - Data on land-use changes for assessing environmental impacts
- **Real Estate Professionals**
 - Are interested in accurate data for property valuation, insights into market trends
- **Legal and Regulatory Bodies, Data Protection Authorities**
 - Ensure compliance with data protection laws, protection of individual privacy rights

Key challenges with respect to agricultural land market data

- **Findable**

- Several [data sources](#) are available, not necessarily findable.
- It can be challenging for those without insider knowledge to locate data that is of interest, given the absence of a common platform, and it's inherent heterogeneity (The data sources differ in terms of purpose, regional coverage, time period, data type, temporal as well as spatial resolution).

- **Accessible**

- Some data are publicly accessible, typically at an aggregated level, Others are labelled as confidential with limited access for defined users; Legal aspects play a major role.

- **Interoperable**

- Data from different sources or fields cannot "talk" to each other yet, Often they lack from a formal, shared, and broadly applicable language.

- **Reusable**

- Is often hindered by various factors, including poor metadata usage licence description, and a weak documentation.
- Do agricultural land market data meet domain-relevant community standards?

Data sources with focus on socio-economics (Examples)

AFiD - Panel Agricultural structure	https://www.forschungsdatenzentrum.de/de/agrar/afid-panel-agrar
ASS – Agricultural structure survey	https://www.destatis.de/
FADN - Farm Accountancy Data Network (FSDN ...Sustainability... as from 2025)	https://agridata.ec.europa.eu/extensions/FarmEconomyFocus/FADNDatabase.html
IACS / InVeKoS - Integrated Administration and Control System	https://agriculture.ec.europa.eu/common-agricultural-policy/financing-cap/assurance-and-audit/managing-payments_en
BORIS - Standard land values	https://www.bodenrichtwerte-boris.de/boris-d
BVVG - Tender results and rental prices	https://www.bvvg.de
Expert committees - Purchase price collections	https://gutachterausschuss.brandenburg.de/gaa/de/gutachterausschuesse/oberer-gutachterausschuss (Brandenburg)
HUB - Long-term field tests	https://www.agrar.huberlin.de/de/institut/einrichtungen/freiland/thyrow/dau_versuch (Thyrow)

Objectives, expected outcomes (1)

- **Linking Agrosystem Data with Socio-economic Information**

- Systematization and evaluation of the current system of data sources on agricultural land markets regarding FAIRagro principles; Identification of opportunities to increase transparency in socio-economic agricultural land market data (→ Report)
- Facilitate access for scientific purposes (→ Development of Technical protocols / Meta analyses)

Example for a meta analysis with focus on IACS / InVeKoS - Integrated Administration and Control System

- Review protocol for identifying and analyzing publications using plot-level IACS data from Austria, Czechia, France, Germany, and Sweden in a systematic map
- 12 academic publications, IACS data from 2005 to 2018

Source: Leonhardt, H., Hüttel, S., Lakes, T., Wesemeyer, M., Wolff, S. (2023): Use Cases of the Integrated Administration and Control System's Plot-Level Data: Protocol and Pilot Analysis for a Systematic Mapping Review. German Journal of Agricultural Economics 72(3/4): 168-184. <https://doi.org/10.30430/gjae.2023.0385>

Results of a meta analysis with regard to IACS data

Datasets combined with IACS data in the sample papers (shorted list, see table 6)	Link(s) to IACS		
	Spatial join	Farm ID	Municipality
Weather data (temperature, precipitation)	X		
Digital elevation model (topographic data)	X		
Regional planning data, Municipality borders	X		
Soil quality data	X		
Land register		X	
Farm Accountancy Data Network (FADN) data		X	X
Agricultural structure survey (ASS)			X

Benefits

- ✓ Available EU wide, yearly collection
- ✓ Reliable and of high quality
- ✓ High level of detail (e.g., crops, plot level, farm level)
- ✓ Information on present/past land use
- ✓ Combines information on land use and farm structure

Limitations

- Not all farmed land/farms included
- Farm IDs cannot be linked to other datasets
- Differences in data setup/collection across the EU
- Differences in data collection/provision over time
- Farm IDs change over time (anonymization)

Suggestions

- Farm IDs
- Crop/livestock management information
- Differentiated/additional use categories
- Information on farmstead locations
- Enable link to other databases such as FADN

Source: Leonhardt, H., Hüttel, S., Lakes, T., Wesemeyer, M., Wolff, S. (2023), p. 179

Objectives, expected outcomes (2)

• Improving Access to Crop Field Trials at HUB

- LASI oversees data from 6 long-term, continuous field trials in Thyrow and 2 additional long-term field trials in Dahlem → Identification of channels for advanced communication and dissemination of these crop field trials

Table 1. General soil parameters at the research station Thyrow, taken from the soil profile next to the experiment field.

Parameter	Topsoil (0–30 cm)	Subsoil (30–60 cm)
Clay (%)	<5	<5–20
Silt (%)	10–14	10–27
Sand (%)	>80	50–80
Bulk Density (g cm^{-3})	~1.6	~1.7
C_{org} (%)	0.4–0.8	<0.02
CEC ¹ ($\text{cmol}_c \text{ kg}^{-1}$)	<5	<5–11
uFC ² (mm)	24	20–66

¹ Cation exchange capacity, ² usable field capacity.

Source: Roß, C. L., Baumecker, M., Ellmer, F., & Kautz, T. (2022). Organic manure increases carbon sequestration far beyond the “4 per 1000 Initiative” goal on a sandy soil in the Thyrow long-term field experiment DIV. 2. Agriculture, 12(2), 170. <https://www.mdpi.com/2077-0472/12/2/170>



Fragmented but Mandatory – A Peculiar Data Landscape

- **16 states, 16 methodologies** – each *Bundesland* collects sales data via its own *Gutachterausschüsse* or statistical office.
- **No national database** – reporting law (*Grunderwerbsanzeige*) guarantees completeness, not standardization.
- **Designed for administration, not research** – primary goals: valuation, taxation, market monitoring.
- **Access hurdles** – only aggregated stats are public; raw transactions require case-by-case agreements.
- **Outcome** – heterogeneity + legal gatekeeping = high entry costs for researchers.

First question of our use case:

- How is farmland data in Germany structured, available and used for research?



Statistics



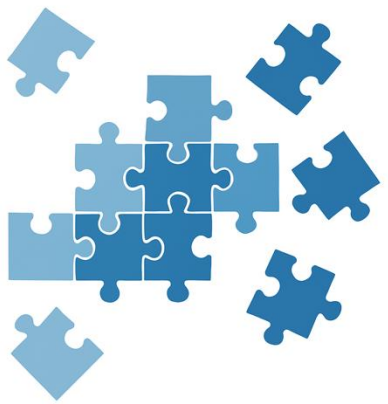
Regulation



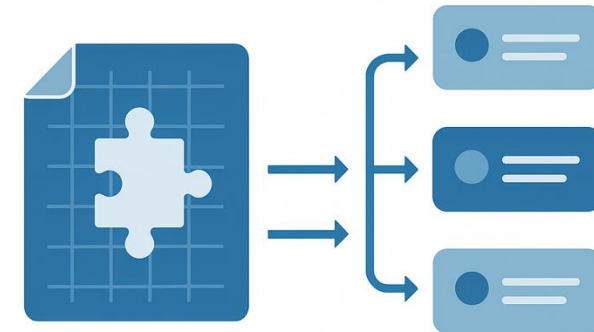
Tax

Why Researchers Struggle: From Fragmented Data to Fragmented Insights

What exists	What researcher needs
Data collected per <i>Bundesland</i>	Harmonized, national-level dataset
Heterogeneous schemas & formats	Standardized metadata and classifications
Focus: administrative use (valuation, tax)	Focus: structural analysis, land market dynamics
Aggregated statistics available	Micro-level transaction data access
Case-by-case access agreements	Transparent and predictable access rules



Legal reporting ensures data collection – but not accessibility, standardization, or analytical value.



Farmland transaction data in Germany

Despite Fragmentation: Research Is Already Happening

1. Use of Available Data

- BVVG auction data (federal land privatization) as a valuable homogeneous dataset
- Selective access to *Gutachterausschuss* data at regional level
- Local data initiatives (e.g., Lower Saxony, Brandenburg, Saxony)

2. Creative Workarounds

- FOI requests or partnerships with local authorities
- Linking public tender data (e.g. from BVVG) with geospatial or ownership info

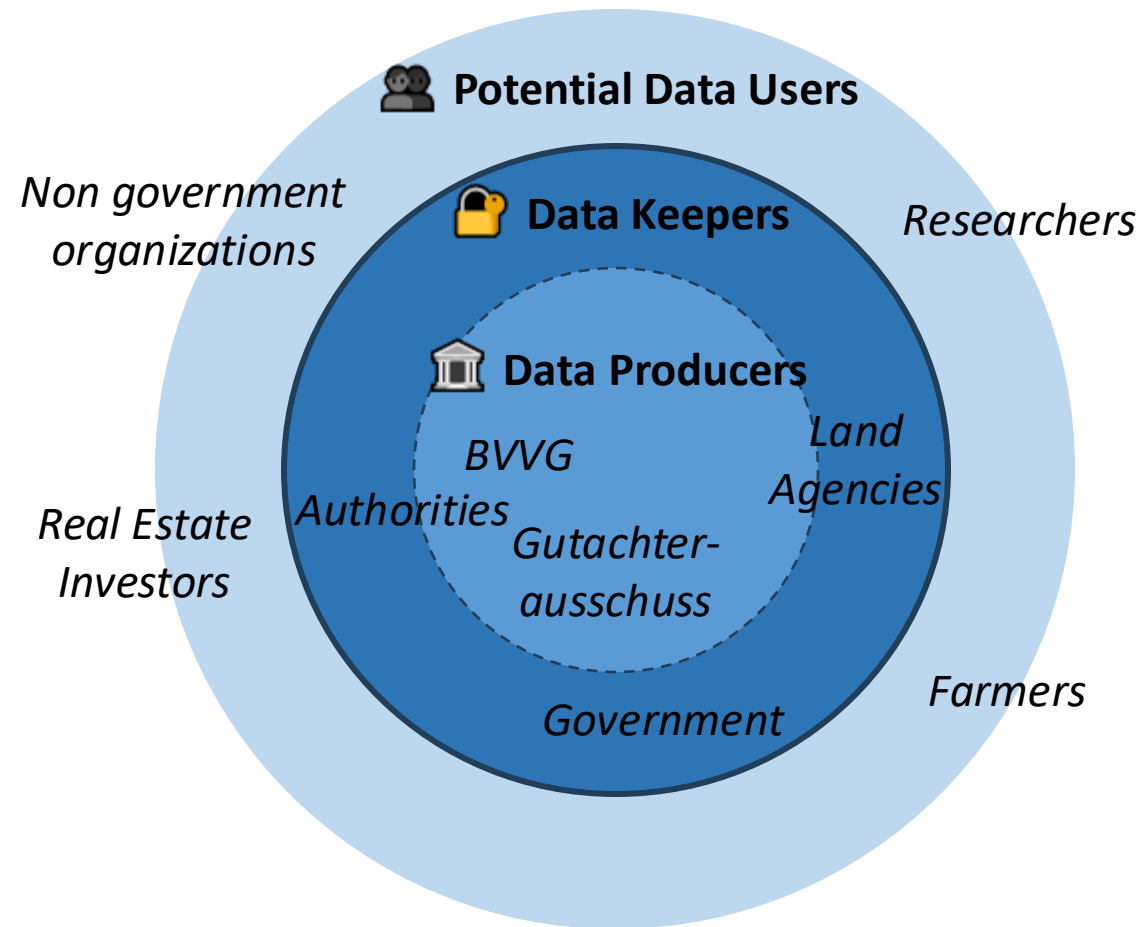
3. Contribution of Our Project

- Inventory of transaction datasets in Germany
- Classification by **structure, access, relevance**
- Assessment of suitability for economic research & policy advice

Even without a central database, studies on land markets in Germany are advancing — creatively and carefully.

Farmland transaction data in Germany

Who Collects What – And For Whom?



Farmland transaction data in Germany

One objective: How is farmland data in Germany structured, accessible, and usable for research?

Core Questions:

- What types of farmland data **exist** (transactions, leases, ownership)?
- How is the data **structured** (spatial resolution, classification systems)?
- Who can **access** it, under what legal or institutional constraints?
- To what extent is the data **usable** for empirical economic research?



Farmland transaction data in Germany

Our Approach: Using AI to create a farmland data inventory for Germany

1. Input: Existing Research & Reports

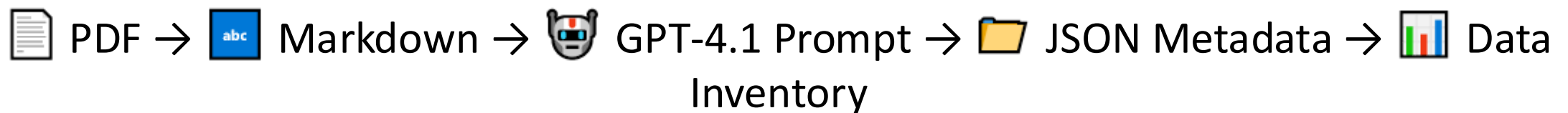
- We gather PDF papers, dissertations, reports using farmland transaction data
- These contain hidden metadata: *data sources, regions, access modes, variables*

2. Process: LLM-Based Metadata Extraction

- Each PDF is converted to **Markdown** format (LLM-native)
- We pass this as input to **GPT-4.1** with strict extraction instructions
- Output is validated **JSON** following a predefined metadata schema

3. Output: Metadata Registry

- Structured inventory of farmland data sources across German states
- Each entry includes: region, provider, years, variables, access conditions
- Basis for FAIR analysis and cross-regional comparisons



Farmland transaction data in Germany

Why an AI-Derived Inventory?

1. Because the data is unstructured

- Most farmland datasets are only *described*, not published.
- Information about them lives in **natural language** inside PDFs: “We use transaction data from Brandenburg 1998–2012 collected by X”.
- No APIs. No standardized repositories. No metadata.
- 🧠 LLMs are optimized for understanding such *free-text descriptions* — exactly where human-driven processes fail or become too slow.

2. Because we need structure from mess

- Our goal is not to extract the full dataset — but to extract a *description of the dataset*.
- Naming the source (e.g. BVVG, Gutachterausschuss), Extracting temporal + spatial coverage, Identifying transaction types (sales, rental, auction)
- Assessing access level, licensing, data quality
- 🛠️ LLMs like GPT-4o allow structured JSON output based on a predefined metadata schema — making the process scalable and standardized.

Farmland transaction data in Germany

Why an AI-Derived Inventory?

3. Because human parsing doesn't scale




- Manual metadata extraction is slow, inconsistent, and subjective.
- Our AI pipeline:
 - Processes 21 papers in seconds, not weeks
 - Enforces consistent field logic across 23 metadata fields
 - Uses AI to perform *semantic deduplication* (e.g., merging datasets that are the same but named differently)

→ This enables us to move from isolated studies to a national **inventory** of farmland data sources.




Traditional data collection fails in fragmented, undocumented environments — AI enables us to extract, harmonize, and consolidate metadata at scale.

From Paper to Inventory: Our Full AI Pipeline



1. PDF to Markdown

-  Input: 21 academic papers
-  Tool: MarkItDown
-  Goal: make LLM-readable




2. LLM Extraction (GPT-4o)

-  Extract 23 fields → JSON metadata
-  Fields: region, variables, access, provider, temporal span
-  Schema-enforced output



3. FAIR Scoring

-  Rule-based and LLM-assisted
-  Highlights poor metadata & access in existing research

4. Semantic Deduplication

-  GPT-4o detects similar/duplicate sources (name, time, region, description)
-  Smart merging to preserve best fields
-  31 raw sources → **27 final entries**

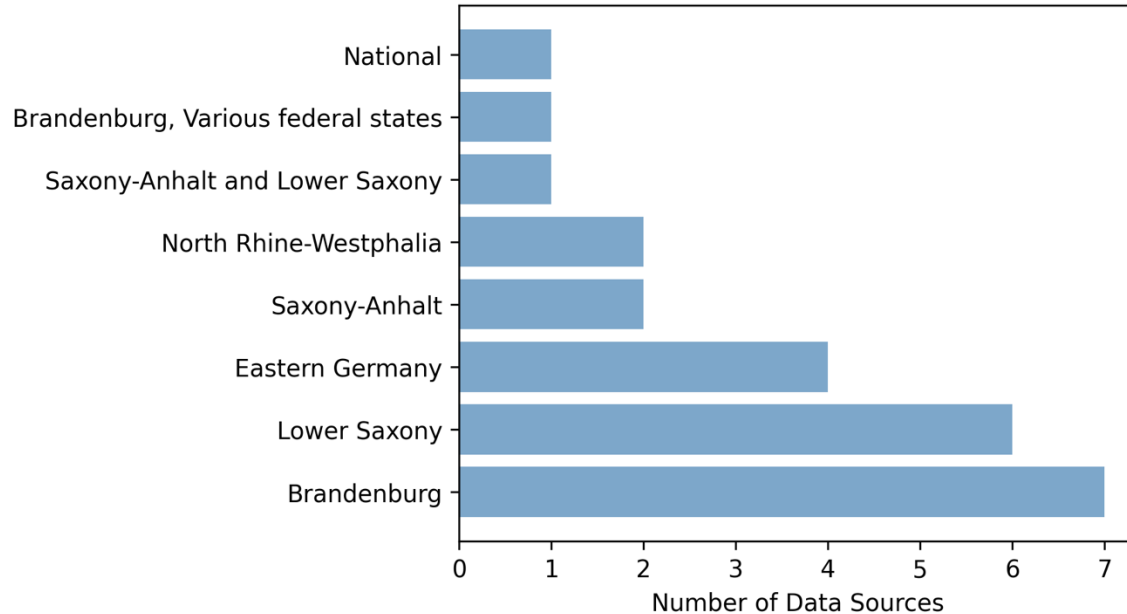
5. Consolidated Registry

-  Output: machine-readable JSON + CSV
-  Ready for integration, visualization, FAIR improvement

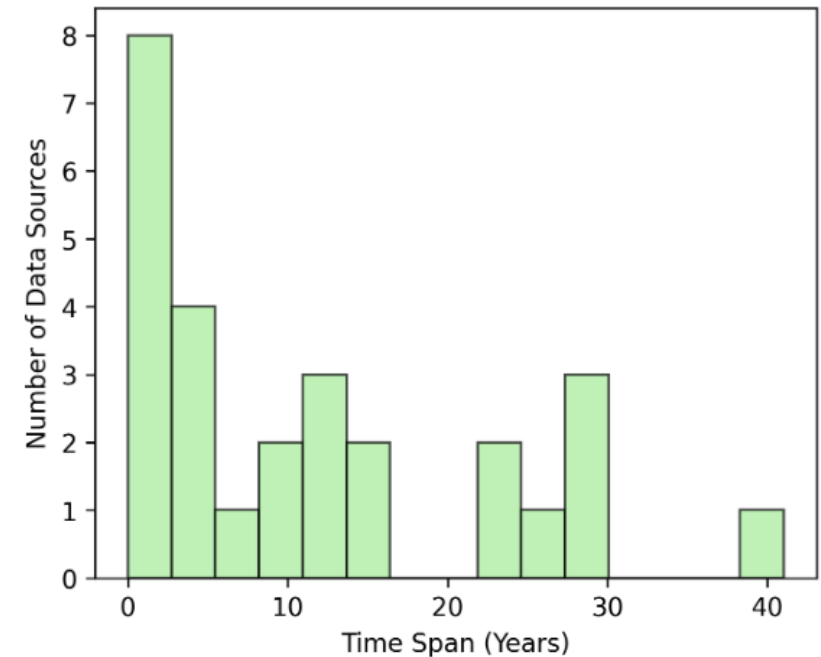
Farmland transaction data in Germany

Key Metrics and Coverage

Top 8 Regions by Data Sources



Distribution of Time Spans



Literature Corpus - *preliminary*

- **21 peer-reviewed papers & reports**
- Domains: land-price analysis, market structure, land-use change
- Time coverage of underlying data: 1975 – 2022

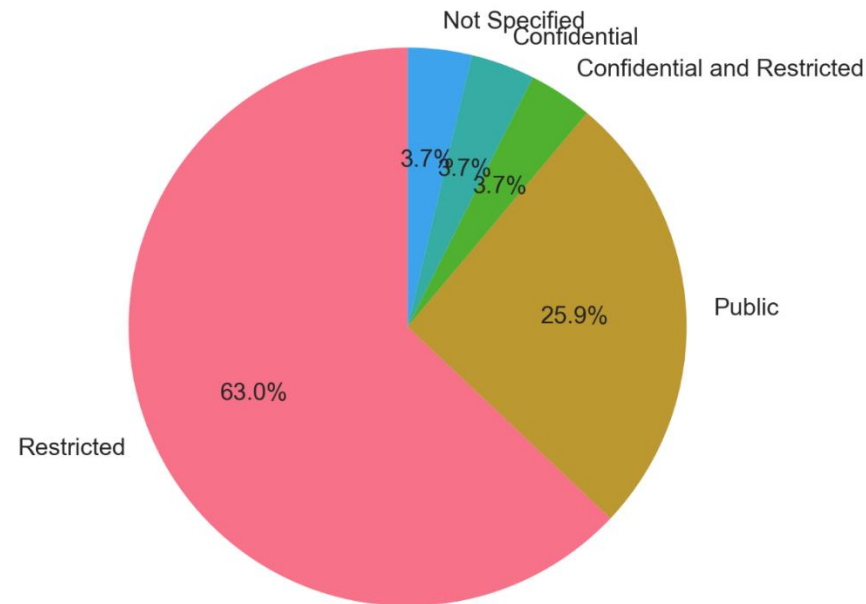
Data types uncovered:

- 45 % administrative records
- 16 % BVVG auction datasets
- 13 % survey/statistical panels
- 13 % official statistics
- 10 % remote-sensing & experimental datasets

Farmland transaction data in Germany

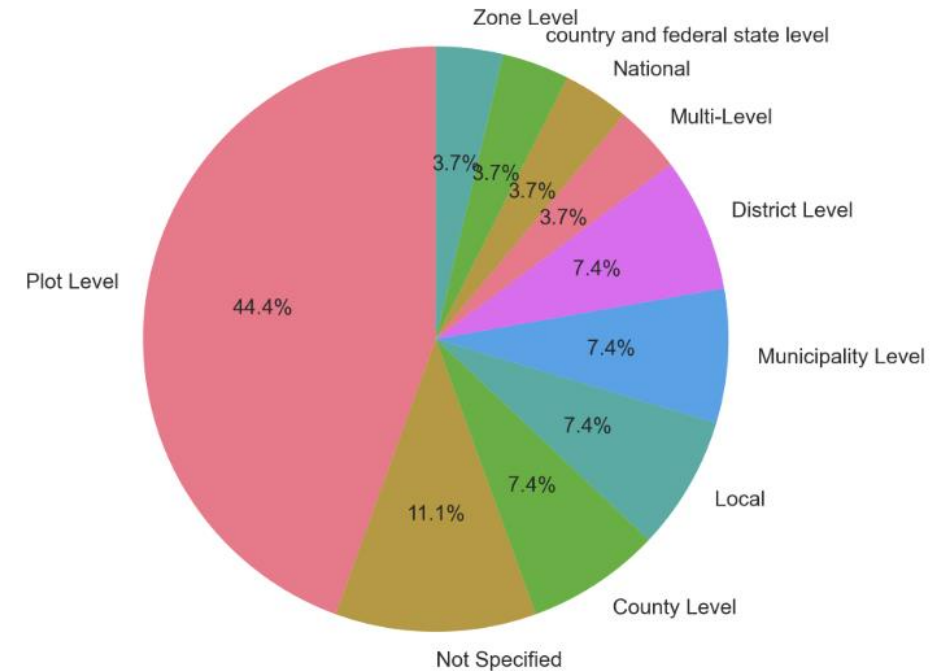
Key Metrics and Coverage

Data Accessibility Distribution



Fine grained transaction data is not publicly available.

Spatial Resolution Distribution



Diverse spatial resolution distribution, depending on data types.

Farmland transaction data in Germany

Next Step #1 – Schema Upgrade & Interoperability

From project-specific JSON to a community-aligned, machine-actionable schema

1. Map to Established Standards

- Cross-walk our 23 fields to **schema.org/Dataset**, **DCAT-AP**, and **DataCite 4.4**.
- Adopt common terms for *spatialCoverage*, *temporalCoverage*, *license*, *isAccessibleForFree*.

2. Link to Domain-Specific Vocabularies

- Align with existing repositories to ensure interoperability.
- Make metadata available so that the already analyzed data sources can be integrated in existing research data repositories.

3. Produce Machine-Readable Artefacts




- Export JSON-LD & CSV metadata templates.

Next Step #2 – Dual FAIR Assessment Engine

Combining deterministic rules with GPT reasoning for richer, auditable scoring

Why we're Upgrading the FAIR Assessment for farmland data

- The FAIR score of all 27 sources is either not assessible or 0.0 — a strong signal that most farmland data is not properly documented or accessible.
- Many deficiencies stem from lack of structured metadata, not data quality itself.
- Traditional FAIR scoring methods are too rigid or too shallow to capture nuance.

Component	Role	What It Does
Rules Engine 	Deterministic	✓ Checks for DOIs, URLs, licenses, metadata fields
LLM Engine (GPT) 	Contextual reasoning	✓ Reads descriptions, infers missing information ✓ Flags ambiguous or implicit reuse statements
Explainability Layer 	Combined output	✓ Every score gets a machine-readable explanation

Looking Ahead – Community Guidelines for Documenting Farmland Data

Better metadata = better research = better policy

1. Develop a Shared Documentation Standard

- Provide templates and examples for farmland data descriptions
- Tailored to *socio-economic use cases*
- Compatible with schema.org, DataCite, and repository requirements

2. Establish Best Practices for Inventory Integration

- Encourage authors to include:
- Time period, region, transaction types
- Access status, usage license, link to raw data (if available)
- Citation of source agency (e.g. Gutachterausschuss, BVVG)
- Include guidance for both historical and ongoing datasets

3. Promote FAIR Thinking in the Community

- Publish a short guideline paper: *“How to Describe Farmland Data for Reuse”*
- Engage with data providers, journals, and research networks (e.g., BonaRes, FAIRagro)
- Provide a simple tool for metadata entry & export

Check Out Our Repository:

Full code and instructions:

https://github.com/lwschm/FAIR_farmland