The first NRW State Research Report: Role, challenges and solutions from the data trustee CRIS.NRW

- Extended Abstract -

Malte Kramer*, Eva Frechen

University of Muenster, Schlossplatz 2, 48149 Münster

Abstract

In 2025, the German federal state North Rhine-Westphalia (NRW) published its first State Research Report, jointly produced by the Ministry of Culture and Science and 32 higher-education institutions (14 universities, 18 universities of applied sciences). CRIS.NRW acted as data trustee. The report is aligned with the Core Data Set for Research (KDSF) standard, avoids inter-institutional rankings, and combines aggregated indicators (Part I) with institution-authored narratives in selected KDSF research fields (Part II). We processed 52 password-protected Excel files across 397 validated sheets and 734,328 rows spanning three reporting years. This extended abstract outlines governance and the trustee model (separation of duties, small-cell policies), pragmatic KDSF profiling, an auditable QA pipeline, a dual perspective on provider and trustee challenges, and an awards case vignette. We conclude with a concise blueprint for schema-based submissions, shared code lists, and an automation roadmap toward APIs and a public portal.

Keywords: Open research information infrastructures; Data trustee governance; Interoperability (KDSF, PIDs); Quality assurance and schema validation; Responsible research evaluation; State-level CRIS reporting

1. Context and contribution

Germany's research and higher education landscape is organized federally: the federal states hold core responsibilities for higher education governance, funding, and reporting. NRW, the most populous German state (~18 million residents), is a state-scale research ecosystem comparable in population to the Netherlands. Building a reusable, state-level evidence base across heterogeneous systems and definitions, without a historical baseline, required a shared semantics and a neutral coordinator. The Core Data Set for Research (KDSF) provided the semantic backbone, consistent with FAIR's emphasis on interoperable, machine-actionable data (Wilkinson et al., 2016). Persistent identifiers (PIDs) such as ORCID and DOI supported entity resolution where available (Haak et al., 2012; Paskin, 2010), and community infrastructures like OpenAlex rely on ROR for organizational affiliation disambiguation, underscoring the importance of open, interoperable metadata (Priem et al., 2022). As a first-of-its-kind effort, target audiences were not sharply defined at the outset (media, policy, leadership, public), affecting indicator granularity, language, and visualization. Our contribution is a concrete, state-scale implementation of a data-trustee model with a reproducible QA pipeline, pragmatic KDSF profiling, and a curated awards taxonomy, accompanied by quantitative process evidence that others can reuse.

2. Governance and the data-trustee model

The rectors' conferences (LRK) mandated a joint working group of institutional leaders, experts, and the state ministry to agree scope and cadence. The ministry financed editorial design and published the report but did not access raw data. CRIS.NRW, as data trustee, integrated submissions, ran QA, produced aggregations, and drafted editorial

* Corresponding author. Tel.: +49 251 83 30357; *E-mail address:* malte.kramer@uni-muenster.de

templates. Contracts for commissioned processing, purpose limitation, and data minimization applied; publication was aggregate-only. A small-cell policy mitigated re-identification risks (Sweeney, 2002). Where person-level checks were technically required for QA, pseudonymization and separation of duties applied; deletion/retention schedules were defined. The model blends bottom-up data-trust principles (Delacroix & Lawrence, 2019) with sector governance, aligning with responsible metrics guidance that prioritizes contextualized, non-ranking communication (Hicks et al., 2015; Waltman, 2016).

3. Indicators and pipeline

KDSF served as the semantic backbone; a pragmatic state profile set indicator definitions, mandatory fields, and permissible aggregation levels. Where KDSF field mapping was not yet feasible across all institutions, DFG/Destatis groupings were used as a transition. Pain points included publication typologies, third-party funding categories and project roles (lead/partner), new state funding lines, patent status/counting, affiliations/organizational units, and time windows. Crosswalks and validation rules operationalized the profile.

Pipeline (auditable, repeatable):
Local CRIS exports → Secure transfer → Schema/rule validation → Normalization → Dedup/entity resolution → Aggregation (small-cell rules) → QA feedback to institutions → Institutional sign-off → Publication

Across 32 institutions and three reporting years we processed 52 files (397 validated sheets; 734,328 rows). Providers reported 35,301 funding projects (raw). PIDs were used where available (ORCID/DOI/ROR), with clear fallbacks to avoid blocking on incomplete PID coverage (Haak et al., 2012; Paskin, 2010; Priem et al., 2022). Checks followed core data-quality dimensions (Pipino et al., 2002), and record-linkage heuristics supported matching and deduplication (Christen, 2012).

4. Dual perspective on challenges

Perspective of Data providers: Two domains required a paradigm shift. First, third-party funding moved from ERP-dominant monetary aggregates to project-based counts with aggregation by subjects, domains, and funders, ideally by programme lines (DFG/EU). Projects are often split over multiple accounts; smaller ones pooled—without CRIS-backed registers, manual compilation is extensive and counting rules diverge. Second, “awards and memberships” were treated as PR lists rather than structured data; submissions mixed prize-like awards with quasi-awards (competitive grants) and varied widely in detail.
Perspective of Trustee: Excel proved practical yet too flexible: added sheets/columns/formatting defeated automation. Many “small” inconsistencies (monetary signs, currency formats, academic titles, ad hoc gender codes) consumed days of normalization because schemas and code lists were not enforced. At times, rule-driven but purpose-misaligned submissions produced low-utility data (projects with no financials; patents without reference year; non-academic events as awards). Uneven counting rules (project vs subproject vs account), inconsistent consortium-role handling, and mixed PID notations undermined comparability.

5. Case vignette: awards and memberships

Institutions reported 2,567 awards/memberships (raw). After deduplication and curation, just over 800 met inclusion criteria (≈31% retention). In the absence of a reference registry, categorization was derived from the data: purposes (excellence, early-career, transfer/innovation, open science), disciplinary scope (including SDG links), eligible audiences (career stage; individual/team/institution; sector; residence), recognized work (single vs corpus; publications; prototypes/patents; transfer; teaching/science communication; open science; engagement/ethics), and quality criteria (peer review, expert judgement). Small-cell rules and narratives replaced any ranking. The resulting schema is reusable for future cycles.

6. Key findings and novelty for the CRIS community

- A neutral data-trustee with separation of duties and small-cell policies reconciles data sovereignty with public reporting at state scale within a federal system.
- KDSF profiling must be pragmatic and iterative; “good enough” profiles, crosswalks, and validation rules enable a first edition and create reusable artifacts.
- Schema discipline is indispensable: locked templates or, better, machine-validated CSV/JSON plus schemas and automated ingestion should replace free-form spreadsheets.
- Versioned code lists are the key to aggregation (programme codes for DFG/EU/federal/state; awards; subject/field systems; organizations via ROR, evidenced by OpenAlex’s use of ROR for affiliations: Priem et al., 2022).
- Purpose-fit acceptance criteria reduce noise and downstream curation costs.
- Early target-audience alignment (personas/use cases) should drive indicator selection, granularity, and visualization.

Novelty: To our knowledge, this is among the first KDSF-based, state-level reports in Germany implemented with a formal data-trustee model and a reproducible QA pipeline, reporting quantitative processing evidence (52 files; 397 sheets; 734,328 rows).

7. Limitations and next steps

First-iteration limitations remain: initially unclear audiences; uneven PID coverage; no award registry; infeasible programme-level aggregation from heterogeneous free text. Next steps include explicit personas/use cases; schema-validated CSV/JSON or APIs with immediate feedback and acceptance only for conformant submissions; shared, versioned code lists (funding programmes, awards, subjects, organizations); and a public portal complementing the triennial report. Privacy protections (aggregation, small-cell rules) and responsible metrics will remain in place (Sweeney, 2002; Hicks et al., 2015; Waltman, 2016).

References

1. Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer. https://doi.org/10.1007/978-3-642-31164-2
2. Delacroix, S., & Lawrence, N. D. (2019). Bottom-up data trusts. International Data Privacy Law, 9(4), 236–252. https://doi.org/10.1093/idpl/ipz014
3. Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID. Learned Publishing, 25(4), 259–264. https://doi.org/10.1087/20120404
4. Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto. Nature, 520, 429–431. https://doi.org/10.1038/520429a
5. Paskin, N. (2010). Digital Object Identifier (DOI) system. In Encyclopedia of Library and Information Sciences (3rd ed.), 1586–1592.
6. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211–218. https://doi.org/10.1145/505248.506010
7. Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully open index of scholarly works, authors, venues, institutions, and concepts. arXiv. https://doi.org/10.48550/arXiv.2205.01833
8. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557–570. https://doi.org/10.1142/S0218488502001648
9. Waltman, L. (2016). A review of the literature on citation impact indicators. Journal of Informetrics, 10(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007
10. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18