Opportunity Title: Scalable and Accurate Entity Resolution in the Presence of Low-Quality Data

Opportunity Reference Code: ICPD-2025-21

Organization	Office of the Director of National Intelligence (ODNI)
Reference Code	ICPD-2025-21
How to Apply	**Create and release your Profile on Zintellect** – Postdoctoral applicants must create an account and complete a profile in the on-line application system. **Please note: your resume/CV may not exceed 3 pages.**

OAK RIDGE INSTITUTE
FOR SCIENCE AND EDUCATION

ORISE GO

The ORISE GO mobile app helps you stay engaged, connected and informed during your ORISE experience – from application, to offer, through your appointment and even as an ORISE alum!

Visit ORISE GO ⬀

GET IT ON Google Play

Download on the App Store

Complete your application – Enter the rest of the information required for the IC Postdoc Program Research Opportunity. The application itself contains detailed instructions for each one of these components: availability, citizenship, transcripts, dissertation abstract, publication and presentation plan, and information about your Research Advisor co-applicant.

Additional information about the IC Postdoctoral Research Fellowship Program is available on the program website located at: https://orise.orau.gov/icpostdoc/index.html.

If you have questions, send an email to ICPostdoc@orau.org. Please include the reference code for this opportunity in your email.

Application Deadline	2/28/2025 6:00:00 PM Eastern Time Zone
Description	**Research Topic Description, including Problem Statement:**

Entity resolution, also known as data matching and record linkage, is the process of identifying records which correspond to the same entity, for example, a distinct person, attribute or product [1].

The National Intelligence Community (NIC) utilizes entity resolution methods to manage datasets from reporting entities and other sources, of varying data quality, consistency and formats. Entity resolution is critical in supporting the data management processes that connect large, disparate data sources to real world entities.

In an increasingly data driven and digital operating environment, entity resolution methods that address emerging data management challenges are required to ensure that financial intelligence production remains comprehensive and actionable for government partners in responding to the evolving threat environment.

There are four main challenges encountered in attempting to resolve entities:

- Scalability of the matching algorithm in terms of overall record processing volume, streaming and number of inputs.11
- Maintaining accuracy of the algorithm as it scales.
- Allowing for incremental history to be included in the processing to further establish the real-world relationships between entities that will improve the linkage accuracy over time.
- Handling for bias in sparse underrepresented information in the source dataset or inputs.

It is therefore important to conduct research on contemporary entity resolution approaches, with the aim of informing enhancements to current applications within financial intelligence processes.

Opportunity Title: Scalable and Accurate Entity Resolution in the Presence of Low-Quality Data
Opportunity Reference Code: ICPD-2025-21

Example Approaches:

The current algorithms tend rely on attribute value similarities to deduce the likelihood of a record pair corresponding to a single entity or different entities with varying historical data quality. Furthermore, where population data is concerned, certain underrepresented data attributes may be overlooked by existing linkage algorithms and or underlying attribute comparison. We would like to see proposals that explore the following approaches:

1. Investigate the use of modern Probabilistic Data Structures (such as Count-min sketch [2] and Bloom filters [3]) to improve the scalability while keeping the high accuracy of existing industrial-grade entity resolution solutions used by the NIC [4, 5].
2. Investigate how to record incremental history comparisons of entity features between each run time session. This includes the comparison of historical entity features to improve or learn from the linkage accuracy of existing entity resolution solutions.

Ideally, applicants should propose to develop protocols or processes that can run both incrementally as well as a batch process and test whether both approaches can produce similar accuracy and efficiency. Novel machine learning approaches are also encouraged but with consideration of ISO standard: ISO/IEC 42001.

Relevance to the Intelligence Community:

Named entity recognition is an ongoing big data challenge for the NIC. Probabilistic methods, particularly those based on recent developments in both academic and commercial applications, stand to address emerging identity management challenges.

References:

- [1] Christen, P. "Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection". Data-Centric Systems and Applications, Springer, Heidelberg (2012)
- [2] Cormode, Graham. "Count-Min Sketch." (2009): 511-516.
- [3] Schnell, Rainer, Tobias Bachteler, and Jörg Reiher. "Privacy-preserving record linkage using Bloom filters." BMC medical informatics and decision making 9 (2009): 1-11.
- [4] Li, Yang, Thilina Ranbaduge, and Kee Siong Ng. "Privacy Technologies for Financial Intelligence." arXiv preprint arXiv:2408.09935 (2024).
- [5] Zhang, Yuhang, Kee Siong Ng, Tania Churchill, and Peter Christen. "Scalable entity resolution using probabilistic signatures on parallel databases." In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 2213-2221. 2018.
- [6] Ranbaduge, Thilina, Peter Christen, and Rainer Schnell. "Large Scale Record Linkage in the Presence of Missing Data." arXiv preprint arXiv:2104.09677 (2021).

- [7] Sarlin, Paul-Edouard, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. "Superglue: Learning feature matching with graph neural networks." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4938-4947. 2020.

Key Words: Record linkage, Scalability, Incremental linkage, ISO/IEC 42001.

Qualifications

Postdoc Eligibility

- U.S. citizens only
- Ph.D. in a relevant field must be completed before beginning the appointment and within five years of the appointment start date
- Proposal must be associated with an accredited U.S. university, college, or U.S. government laboratory
- Eligible candidates may only receive one award from the IC Postdoctoral Research Fellowship Program

Research Advisor Eligibility

- Must be an employee of an accredited U.S. university, college or U.S. government laboratory
- Are not required to be U.S. citizens

Point of Contact

Keri Tarwater

Eligibility Requirements

- **Citizenship:** U.S. Citizen Only
- **Degree:** Doctoral Degree.
- **Discipline(s):**
  - **Chemistry and Materials Sciences** (12 👁)
  - **Communications and Graphics Design** (3 👁)
  - **Computer, Information, and Data Sciences** (17 👁)
  - **Earth and Geosciences** (21 👁)
  - **Engineering** (27 👁)
  - **Environmental and Marine Sciences** (14 👁)
  - **Life Health and Medical Sciences** (45 👁)
  - **Mathematics and Statistics** (11 👁)
  - **Other Non-Science & Engineering** (2 👁)
  - **Physics** (16 👁)
  - **Science & Engineering-related** (1 👁)
  - **Social and Behavioral Sciences** (30 👁)