NAR Genomics and Bioinformatics, 2026, **8**, lqag037
https://doi.org/10.1093/nargab/lqag037
Opinion Article

A birth certificate for data to improve findability, accountability, and traceability

Rongbin Li[1], Avisha Das[1], Yuntao Yang[1], Zhao Li[1], Na Hong[2], Hua Xu[2], Maryann Martone[3], W. Jim Zheng[1,*]

[1]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX 77030, United States
[2]Section of Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, CT 06520, United States
[3]Professor Emeritus, Department of Neuroscience, University of California, San Diego, San Diego, CA 92093-0608, United States

*To whom correspondence should be addressed. Email: wenjin.j.zheng@uth.tmc.edu

Abstract

Data quality is the foundational to scientific research and the rapid advancement of artificial intelligence. Ensuring data quality, provenance, and reproducibility requires robust mechanisms for traceability and accountability from the moment data are created. We propose the concept of a Data Birth Certificate, a universal framework for identifying research data at creation with built-in provenance information, including time, location, and data creator. Unlike existing identifiers assigned at deposition, a Data Birth Certificate establishes immutable, origin-centered traceability that complements established principles of Findability, Accessibility, Interoperability, and Reusability (FAIR). By capturing essential metadata at data generation, Data Birth Certificates support reliable data tracking, accountability, and downstream information management without constraining how data are stored or reused. This perspective outlines the conceptual framework, distinguishes it from existing identifier systems, and discusses its potential role in strengthening research reproducibility and data stewardship across scientific domains.

Introduction

High-quality data plays a critical role in the current AI revolution, serving as the foundation to train robust and reliable models [1]. Landmark resources such as ImageNet [2] have shown how large, well-annotated datasets can drive breakthroughs in artificial intelligence. However, assembling such resources demands substantial time and effort to track down data—data must be located, verified for quality, standardized for consistency, and thoroughly annotated for model training.

The FAIR Data Principles—Findable, Accessible, Interoperable, and Reusable—have helped address these challenges by significantly improving data traceability. Established persistent identifier (PID) systems such as the digital object identifier (DOI) [3] and the persistent identifier consortium for eResearch (ePIC) [4] provide systematic support for FAIR data management. Similarly, the international generic sample number (IGSN) [5] sets a strong standard for incorporating interpretable local metadata into data identification systems.

Yet, the importance of data traceability extends beyond this utilitarian perspective. Two critical aspects remain underemphasized:

1. Accountability and reachability—ensuring that data generators remain identifiable and contactable for clarification or follow-up. This is particularly important for scientific research.
2. Context and provenance—capturing the environment, experimental conditions, and processes under which data were generated, which is essential for detecting inconsistencies or assessing quality.

To bridge these gaps, we propose a novel concept: the creation of a "data birth certificate"—a system designed to enhance traceability, accountability, and contextual transparency, thereby strengthening data integrity and reproducibility in the era of AI.

The proposed concept of a "data birth certificate" can be understood through an analogy to personal identification systems. In this analogy, existing PID systems for data, such as DOIs [3] or ePICs, function much like a social security number—they provide a unique, standardized identifier that enables tracking, management, and utility across databases and systems. These identifiers are essential for ensuring data FAIR within the broader research ecosystem.

In contrast, a data birth certificate would serve a role similar to an individual's birth certificate, documenting the origin, context, and provenance of the data at the moment of its creation. It would capture key details such as the data's creator, location, experimental conditions, and instruments used—information critical for establishing authenticity, accountability, and reproducibility.

Both identifiers are indispensable but serve distinct and complementary purposes: the PID facilitates data utility and trackability, while the data birth certificate ensures data traceability, accountability, contextual transparency, and provenance, together forming a more complete foundation for data integrity and reliability in scientific research. While PIDs, for

Received: October 4, 2025. **Revised:** February 11, 2026. **Accepted:** March 5, 2026

Table 1. A comparison of Data Birth Certificate and PID

Property	Data Birth Certificate	PID	Examples and notes
Metaphor	Birth Certificate	Social security number	Analogous to how birth certificates and SSNs serve distinct but complementary roles.
Time of creation	Generated at the moment data are generated	Assigned when data are deposited into a repository or tracking system	FATID vs. GEO accession ID.
Uniqueness	Globally unique for each dataset, derived from creator, location, time, and project	Unique within a given system; multiple PIDs may exist across repositories	A gene expression dataset may have a publication with DOI and a GEO ID, requiring cross-referencing.
Consistency	Immutable once created	May change due to versioning, merging, correction, or deprecation	Like a birth certificate, Data Birth Certificate never changes; Some repository IDs may evolve.
Creator information	Explicitly captured and required	Often implicit or absent	Data creators and depositors may differ, especially in large collaborations.
Timestamp	Records when data are generated	Records when data are deposited	Delays between creation and deposition can obscure provenance.
Presence	Exists for every dataset, including failed or discarded experiments	Exists only for deposited datasets	Failed experiments may lack PIDs but still require traceability.
Traceability	Precise trace-back to creator, location, time, and project	Typically traceable to a group or repository	Detailed provenance often requires additional investigation for PIDs.
Primary function	Accountability and provenance (looking backward)	Data management, access, and reuse (looking forward)	The two systems serve complementary purposes. However, information in Data Birth Certificate could be used as important metadata or data for the latter.
System integration	Independent of downstream data management systems	Integrated into repository and information system design	PIDs often act as primary keys in databases; birth certificates do not assume system structure.
Scope of use	Focused on data creation context at the beginning of a data lifecycle	Focused on data lifecycle and reuse	Birth certificates fix origin; PIDs evolve with data use and management.

example, lack critical metadata, the birth certificate can fill this gap by providing comprehensive details about the data's origin, including timestamps, project information, and generator identity, right at the time of creation (Table 1).

In biomedical research, the concept of a data birth certificate is particularly relevant to the growing concerns over data quality and reproducibility [6]. While significant attention has been devoted to improving experimental techniques and analytical methodologies [7], the traceability and accountability of data—fundamental pillars of credible and reproducible science—remain insufficiently addressed. A well-implemented Data Birth Certificate, embedding essential details such as the time, location, how the data is generated, and identity of the data generator at the moment of creation, can help resolve common challenges researchers face, including missing documentation, incomplete metadata, misplaced samples, and gaps in experimental context. By ensuring that key provenance information is recorded at the point of data generation, this initiative shifts the focus from reactive data correction to proactive data integrity, establishing a culture of accuracy, transparency, and accountability that strengthens the foundation of biomedical research.

Implementing Data Birth Certificate and deploy it in practical use requires some careful consideration. For example, like a birth certificate, elements of the Data Birth Certificate should remain constant or fall within a predefined value set for consistency and stability. To maximize stability, utility, and interoperability, the Data Birth Certificate should integrate seamlessly with established identification systems such as ORCID [8], uniform resource name (URN) [9], uniform resource locator (URL) [10], or DOI [3]. For example, in the biomedical domain, we can identify a data creator by their ORCID [8] as the unique identifier (see example below). Doing so can uniquely identify each data creator based on an established single source, avoiding redundancy resulted from multiple ID systems that could lead to data creator untraceable.

The creation and maintenance of Data Birth Certificates should be automated to minimize burden on data creators and reduce human error. We envision a specialized app or device—similar to an inventory scanner with barcode printing and wireless connectivity—that can automatically generate and assign a birth certificate at the point of data creation and log it into a central registry. Incorporating large language models and AI agents [11] would further enhance usability by enabling verbal interaction through a chatbot and supporting tasks such as protocol retrieval, experimental metadata capture, and automatic creator assignment. With

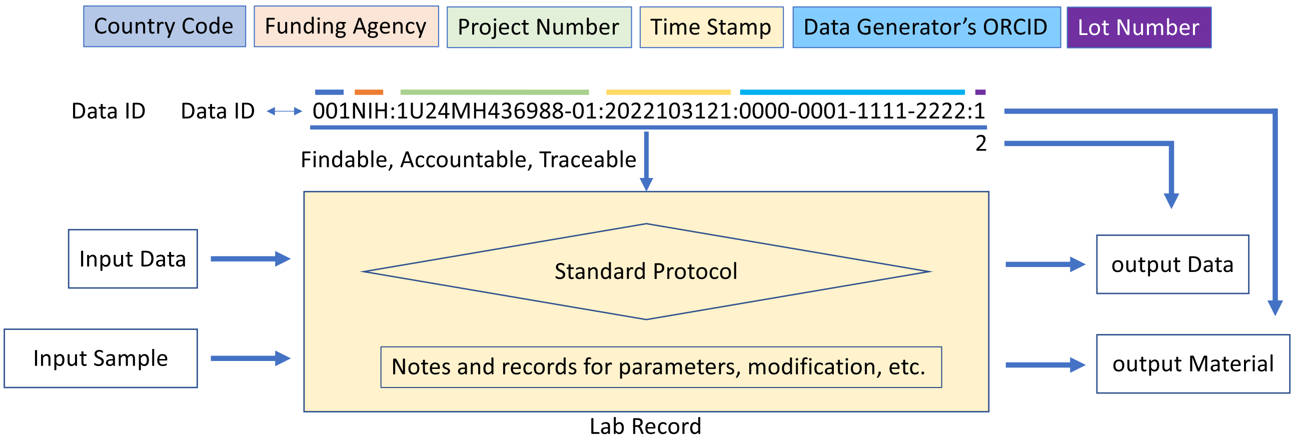

Figure 1. FATID system. FATID components and how they are assembled. Lot #1 is used for output material and Lot #2 is used for output data.

recent advances, AI agents could also assist in experimental design, helping automate the data-generation process from the outset.

Additional technologies such as QR codes, Web APIs, and novel ID-embedding techniques can strengthen reliability and reduce manual steps. When integrated with standard lab protocols and careful record-keeping, this automated system will ensure effective data findability, accountability, and traceability—core elements of reproducible biomedical research and FAIR (Findable, Accessible, Interoperable, and Reusable) data stewardship [12].

Here, we use an example, FATID, for the domain of biomedical research to illustrate a data birth certificate (Fig. 1). FATID, a universal identification (ID) system with built-in **F**indability, **A**ccountability, and **T**raceability, aligns with FAIR principles [12]. This system encodes dataset IDs, integrating existing systems like ORCID, ensuring uniqueness for data generated worldwide (Fig. 1). FATID serve as an example illustrating enhanced transparency and traceability for biomedical research, but there could be many alternative implementations for other settings.

The first field of FATID comprises a three-digit unique international country telephone code, indicating the country of data origin. This is followed by the second field, featuring a three-letter abbreviation representing the country-specific funding agency supporting the research project [13]. A distinct code, such as ZZZ, may be utilized to signify a nonfunded project.

The third field represents a unique grant number that supports data generation. Each grant (e.g. from NIH) is assigned a distinctive number for identifying crucial information about the funded project. Unfunded projects may utilize a personal identifier (e.g. ORCID) of the principal investigator (PI).

The fourth field is a timestamp indicating when the data was created. It serves multiple purposes: (i) indicating the data generation time, (ii) locating lab records for details on data generation, and (iii) checking for errors, such as ensuring the dataset timestamp is not older than those of the data and experimental materials used.

The fifth field is a consistent identifier for the data generator, unaffected by changes in affiliation or location. Privacy concerns are mitigated as data generators play a role like authors in publications. In the academic context, this could be the ORCID of the data generator.

The last field is the lot number, distinguishing different data or material generated from the same experiment. For instance, if an experiment yields both data and samples, they can be labeled with distinct lot numbers, allowing seamless integration of physical objects (e.g. lab samples) and digital objects (e.g. data).

Beyond the FATID example, a Data Birth Certificate can be implemented in various forms depending on the domain, scope, and data-generation process. As long as the critical components are properly defined and recorded, each implementation remains independent and generalizable.

Implementing a Data Birth Certificate requires careful consideration. While capturing detailed data provenance is essential, manual documentation for every dataset is burdensome and prone to error. Automation is therefore critical to reduce workload, improve consistency, and minimize human error at the point of data creation.

A second key challenge is clearly defining the role of the Data Birth Certificate. As summarized in Table 1, the birth certificate records provenance at the moment data are generated and should remain immutable thereafter. Issues such as versioning, data evolution, and relationships among datasets produced by multiple stakeholders should be handled at the data management and implementation level. In this framework, the Data Birth Certificate provides stable, foundational provenance but does not track downstream usage or modifications. These functions are better served by information management systems—such as local LIMS or community repositories—which can use PIDs and leverage birth certificate metadata for lifecycle management without compromising provenance integrity.

It is time to rethink data stewardship from the moment data are created. By engaging the research community in defining and refining the concept of Data Birth Certificates, we can establish a shared foundation for transparency, accountability, and trust in scientific data. Such a foundation will enable thoughtful, domain-specific implementation strategies, including integration with existing identifier systems and data management infrastructures. Together, these efforts can strengthen the reliability of scientific records and support more robust, reproducible discovery in the years ahead.

Acknowledgements

Authors contribution: R.L., M.M., and W.J.Z. conceived the idea, with significant input from N.H., H.X., Z.L., and Y.Y. R.L. and A.D. developed the methodology and visualization. R.L. and W.J.Z. wrote the original draft. All the authors contributed to the reviewing, editing, and finalizing the manuscript. R.L., A.D., and W.J.Z. acquired funding for this research, and W.J.Z. provided overall supervision.

Conflict of interest

None declared.

Funding

This work is partly supported by the National Institutes of Health (NIH) through grants 1UM1TR004906-01, 1R01AG066749, 1U24MH130988-01, and 1R01MH142410-01, Department of Defense W81XWH-22-1-0164, and the Cancer Prevention and Research Institute of Texas through grant RP170668 (WJZ). Rongbin Li is supported by a training fellowship from the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics & Data Science (T15LM007093). Avisha Das is a CPRIT Postdoctoral Fellow in the Biomedical Informatics, Genomics and Translational Cancer Research Training Program (BIG-TCR) funded by Cancer Prevention & Research Institute of Texas (CPRIT RP210045).

Data availability

No new data were generated or analyzed in support of this research.

References

1. Ching T, Himmelstein DS, Beaulieu-Jones BK *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018;15:20170387. https://doi.org/10.1098/rsif.2017.0387
2. Deng J, Dong W, Socher R *et al.* ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami: IEEE, 2009, 248–55. https://doi.org/10.1109/CVPR.2009.5206848
3. Paskin N. Digital object identifier (DOI®) system. *Encycl Libr Inf Sci* 2010;3:1586–92.
4. PID Consortium. ePIC persistent identifiers for eResearch, *Verfügbar unter*. 2020. https://www.pidconsortium.net (12 November 2025, date last accessed).
5. Ramdeen S, Lehnert K, Klump J *et al.* International Generic Sample Number. *Encyclopedia of Mathematical Geosciences*. Cham: Springer, 2023, 656–60.
6. Oza A. Reproducibility trial: 246 biologists get different results from same data sets. *Nature* 2023;622:677–8. https://doi.org/10.1038/d41586-023-03177-1
7. Improving the reproducibility of metabolic research. *Nat Metab* 2022;4:1085. https://doi.org/10.1038/s42255-022-00653-2
8. Bohannon J, Doran K. Introducing ORCID. *Science* 2017;356: 691–692. https://doi.org/10.1126/science.356.6339.691
9. Sollins K, Masinter L. RFC1737: functional requirements for uniform resource names. 1994. https://datatracker.ietf.org/doc/html/rfc1737 (12 November 2025, date last accessed).
10. Berners -Lee T, Masinter L, McCahill M. Uniform resource locators (URL). 1994, No. rfc1738. 1994.
11. Russell SJ, Norvig P. *Artificial Intelligence: A Modern Approach*, 2nd edn. Hoboken: Pearson, 2021.
12. Wilkinson MD, Dumontier M, Aalbersberg IJ *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:1–9. https://doi.org/10.1038/sdata.2016.18
13. GitHub. Funding Agencies. https://github.com/MachinicleInf3/FATID (12 November 2025, last accessed).

Received: October 4, 2025. **Revised:** February 11, 2026. **Accepted:** March 5, 2026