Integrating ARK Persistent Identifiers into Research Data Infrastructure

Fajar Hakka
Teuku Umar University

Abstract: *The integration of Archival Resource Key persistent identifiers into research data infrastructure addresses the growing need for durable, actionable, and interoperable links to digital scholarly resources. ARKs provide a flexible and decentralized identifier framework that supports long-term access, precise citation, and reliable linking across heterogeneous collections. Their adaptability to both machine-readable and human-readable representations makes them suitable for a wide range of research outputs, including evolving datasets and non-traditional materials. By embedding ARKs into repository workflows, institutions can enhance metadata quality, strengthen interoperability, and align with open science mandates while retaining local control over identifier namespaces. The adoption of ARKs complements other persistent identifier systems, such as DOIs, creating a more resilient and inclusive scholarly communication ecosystem. While integration presents technical, policy, and sustainability challenges, the benefits for discovery, preservation, and reproducibility underscore ARKs' potential to play a central role in global research data stewardship.*
Keywords: *ARK persistent identifiers, Research data infrastructure, Digital preservation, Metadata interoperability, Scholarly communication, Open science*

The proliferation of digital research outputs over the past two decades has transformed the way scholarly communities create, store, share, and preserve knowledge. From datasets and code repositories to multimedia materials and interactive models, research artifacts are increasingly born-digital, and their continued accessibility depends on robust systems for identification and management. One of the most significant challenges in this evolving digital landscape is ensuring that research objects remain findable, accessible, interoperable, and reusable over time, regardless of changes in technology, institutional priorities, or hosting platforms. Persistent identifiers have emerged as a cornerstone of modern research data infrastructure, offering a mechanism for assigning stable, actionable, and unique references to digital objects. Among these, the Archival Resource Key system has gained recognition for its flexibility, longevity, and compatibility with diverse repository environments.

The ARK framework is designed to meet the enduring needs of digital preservation while accommodating the realities of technological change. Unlike identifiers that are tied exclusively to centralized governance or narrow domains, ARKs offer a decentralized approach that empowers institutions to maintain control over their identifier namespaces while adhering to a common syntax and resolution mechanism. The result is a system that is both globally recognizable and locally sustainable, providing a stable bridge between the transient world of web addresses and the enduring requirements of scholarly citation. As research data infrastructure becomes more complex and interconnected, integrating ARK persistent identifiers into these systems offers significant advantages in terms of metadata quality, long-term access, interoperability, and trustworthiness.

The integration of ARKs into research data systems begins with an understanding of their fundamental design principles. At their core, ARKs are HTTP-based identifiers that can be resolved to a specific digital object, its metadata, or related contextual information. This design makes them immediately usable within the existing web ecosystem while also supporting the resolution of

multiple representations of the same resource. For example, an ARK can resolve to a machine-readable metadata record for automated harvesting, a human-readable descriptive page for researchers, or a versioned dataset for direct download. This capacity to deliver context-appropriate representations aligns with the evolving practices of data dissemination in research communities, where users may have very different needs depending on their role, discipline, or analytical objectives.

One of the strengths of ARKs in research data infrastructure is their suitability for heterogeneous collections. Research outputs do not exist in isolation; they are often linked to related datasets, publications, software tools, protocols, and institutional records. ARKs can serve as stable anchors for each component of this interconnected web, enabling precise citation, reliable linking, and reproducible workflows. In disciplines such as archaeology, climate science, or genomics, where datasets may be updated, corrected, or augmented over many years, ARKs provide a means to reference both the evolving resource as a whole and its specific historical states. This versioning capability is essential for ensuring that future researchers can accurately reconstruct the conditions under which prior analyses were conducted.

The integration of ARKs into research data infrastructure also intersects with broader developments in metadata standards and interoperability frameworks. Research repositories increasingly operate in federated networks, exchanging information through standardized protocols such as OAI-PMH, ResourceSync, and Schema.org markup. ARKs enhance these interactions by providing unambiguous identifiers that can be reliably matched across systems, reducing the risk of duplication, fragmentation, or misattribution. Furthermore, the commitment to maintaining ARKs over the long term supports the trustworthiness of metadata records, which is a critical factor in data discovery and reuse. When researchers, funders, or policy-makers encounter an ARK, they can be confident that it represents a stable and curated link to the underlying resource.

Another dimension of ARK integration concerns sustainability and governance. Persistent identifiers are only as reliable as the infrastructure and institutional commitments that support them. The ARK system addresses this by allowing organizations to manage their own identifier namespaces while adhering to an open, community-driven standard. This distributed governance model reduces dependency on single points of failure and enables flexibility in adapting to evolving institutional capabilities. For research data infrastructure, this means that repositories can adopt ARKs without surrendering control over their long-term management, while still benefiting from the global recognizability and interoperability that PIDs provide.

In practical terms, integrating ARKs into research data systems involves aligning repository workflows with identifier assignment, resolution, and maintenance processes. This begins at the point of resource creation or ingestion, where an ARK can be minted and associated with the object in both internal records and public metadata. The persistence commitment inherent in the ARK framework requires that institutions ensure the ongoing availability of both the object and its metadata, even if the object itself is deprecated or relocated. This may involve the use of redirection services, mirror sites, or archival storage systems to maintain the integrity of resolution links over time.

The relationship between ARKs and other PIDs, such as DOIs, is often complementary rather than competitive. While DOIs have strong adoption in scholarly publishing and formal data citation, ARKs offer a broader scope that includes unpublished materials, ephemeral resources, and non-traditional research outputs. In many cases, the same object may be assigned both a DOI for citation purposes and an ARK for integration into a broader preservation or archival strategy. The coexistence of multiple PIDs enhances the resilience and accessibility of research objects, ensuring that they can be discovered and referenced through multiple channels.

The benefits of ARK integration extend beyond the immediate concerns of technical infrastructure to influence the cultural and policy dimensions of research data management. Funding agencies and scholarly societies increasingly mandate the use of PIDs to promote transparency, reproducibility, and open science. By adopting ARKs, research institutions can demonstrate compliance with these mandates while also positioning themselves as active participants in a global network of sustainable digital stewardship. The visibility and traceability afforded by ARKs can also enhance the reputation of institutions, as their contributions to the scholarly record become more easily discoverable and citable.

There are, however, challenges to be addressed in the process of ARK adoption. Technical integration requires investment in software development, metadata curation, and staff training. The governance of ARK namespaces demands clear institutional policies, including decisions about versioning, deprecation, and access control. Perhaps most importantly, the long-term success of ARKs in research data infrastructure depends on sustained commitment to preservation and resolution services, even in the face of budgetary constraints or shifting priorities. These challenges are not unique to ARKs but are common to all forms of persistent identification, underscoring the importance of embedding PID strategies within broader institutional planning for digital preservation.

Emerging trends in research data infrastructure point to opportunities for deepening the role of ARKs in supporting new forms of scholarly communication. The growth of linked data and semantic web technologies creates pathways for ARKs to function as stable nodes in interconnected knowledge graphs, linking datasets to publications, funding sources, researchers, and institutions in machine-readable ways. In computational research, ARKs can be embedded within workflows and scripts to ensure that all data inputs and outputs are traceable and reproducible. In community-driven repositories, ARKs can serve as enduring references for collaborative datasets that may evolve organically over time, preserving their provenance and integrity.

The future integration of ARKs may also be shaped by advances in automated metadata generation, where artificial intelligence tools can assist in describing and categorizing resources at scale. This could accelerate the process of assigning and updating ARKs, particularly for large and dynamic collections. Additionally, the adoption of ARKs in emerging fields such as virtual reality archives, 3D model repositories, and complex multimedia datasets may broaden their application beyond traditional text-based or tabular research outputs. By supporting diverse resource types, ARKs can remain relevant in a research environment that continues to diversify in format and complexity.

In conclusion, integrating ARK persistent identifiers into research data infrastructure is a strategic step toward ensuring the durability, discoverability, and interoperability of scholarly resources. The strengths of the ARK system - its flexibility, decentralization, and compatibility with existing web protocols - make it a valuable complement to other persistent identifier systems. While challenges remain in terms of technical implementation and institutional commitment, the benefits of ARKs in supporting long-term access, reliable citation, and interdisciplinary collaboration are compelling. As research data infrastructures continue to evolve, ARKs are well-positioned to play a central role in the preservation and accessibility of the digital scholarly record, ensuring that the knowledge created today remains available and meaningful for generations to come.

References

1. Abrams, S., Kunze, J., & Loy, D. (2020). The ARK identifier scheme. California Digital Library. https://arks.org/about/

2. Ball, A., & Duke, M. (2015). How to cite datasets and link to publications. Digital Curation Centre. https://doi.org/10.18452/16357

3. Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., & Clark, T. (2019). A data citation roadmap for scholarly data repositories. Scientific Data, 6(28). https://doi.org/10.1038/s41597-019-0031-8
4. Kunze, J. A., & Rodgers, R. (2001). The ARK identifier scheme. Internet Engineering Task Force (IETF) Draft. https://arks.org/ark-spec/
5. Lannom, L., Koureas, D., & Hardisty, A. R. (2020). FAIR data and services in biodiversity science and geoscience. Data Intelligence, 2(1–2), 122–130. https://doi.org/10.1162/dint_a_00034
6. Meadows, A., Haak, L. L., & Brown, J. (2019). Persistent identifiers: The building blocks of the research information infrastructure. Insights, 32(1), 9. https://doi.org/10.1629/uksg.457
7. National Information Standards Organization. (2021). Persistent identifiers in scholarly communications. NISO RP-31-2021. https://doi.org/10.3789/niso-rp-31-2021
8. Parsons, M. A., Duerr, R., & Minster, J. B. (2010). Data citation and peer review. Eos, Transactions American Geophysical Union, 91(34), 297–298. https://doi.org/10.1029/2010EO340001
9. Paskin, N. (2009). Digital object identifier (DOI) system. In Encyclopedia of Library and Information Sciences (3rd ed., pp. 1586–1592). CRC Press. https://doi.org/10.1081/E-ELIS3-120044418
10. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18