Decentralized Resource Identifiers in the Research Landscape

Carly Huitema, University of Waterloo and University of Guelph, Guelph, ON, Canada

Robert Mitwicki, Human Colossus Foundation, Geneva, Switzerland

Dave McKay, Cybersecurity Research Lab at Ted Rogers School of Management at Ryerson University, Toronto, ON, Canada

Wenjing Chu, Futurewei Technologies, Santa Clara, CA, USA

Dian Ross, Blockchain@UBC, Vancouver, BC, Canada

Keywords: resource identifiers, research identifiers, ecosystem, DIDs, decentralized, verifiable credentials, PIDs, persistent identifiers

Short Description: A new ecosystem of decentralized research identifiers is proposed that offers benefits of decentralization, interoperability, security, and scalability compared to current implementations.

Executive Summary: Research objects (e.g., authors, samples, equipment, and datasets) need digital identifiers that are persistent, unique, and globally resolvable to enable a host of benefits in the research ecosystem. The current organization of these digital identifiers means that for each new class of identifier there is de novo creation of the requisite supporting identifier system. We propose to outline an ecosystem of Decentralized Resource Identifiers (DRIs), compatible with existing identifiers but offering numerous benefits including decentralization, interoperability, and security. The DRI ecosystem would consist of the technology infrastructure and related tools such as guiding principles, governance examples, semantics, training, documentation.

Current Research Landscape

The importance of digital identifiers for research objects that are persistent, unique, and globally resolvable is increasing 1 but creating new systems of identifiers remains challenging. Resource identifiers are currently used by the research community for multiple resources including documents (DOI), researchers (ORCID), research organizations (ROR), datasets, and much more. However, the creation of each additional type of identifier requires the de novo creation of the requisite supporting identifier system.

Identifiers are important and can ensure accurate credit, recognition, resource tracking, ease of administrative and reporting requirements, discovery, trustworthiness, ethics, reproducibility, auditability, integrity through hashing, and more. Identifiers could be extended far beyond their current use to include many different types of research assets including equipment, grants, collections such as culture collections or strain constructs, metagenomic libraries, code snippets,

research methodologies, teams of researchers, biobanks, samples, conferences, etc. Having a suite of different identifier systems available will be useful to everyone within the research ecosystem.

Existing identifier systems are useful but come with a set of challenges that make them difficult to implement for new use cases. Challenges include the following.

- **Centralization**- Most of the existing identifiers require a centralized registry which all parties in the ecosystem need to trust. This kind of system is prone to security breaches and can be hard to scale. It introduces artificial borders and limitations and creates gatekeepers controlling the flow of users and information, potentially at significant costs to the research community.
- **Lack of interoperability** - A lack of interoperability (e.g., heterogeneous data format requirements) results in a multitude of identifiers across systems all representing the same object. This redundancy introduces additional complexity and makes it harder to maintain systems, track objects, and maintain an agreed upon state of up-to-date information.
- **Fragmentation** - Different types of identifier systems have designed and implemented their own use case specific systems,which increases the costs of development, the cost of maintenance, and introduces interoperability challenges. Verifiability and trust - Existing identifiers do not provide built-in verifiability, which means users must rely on a centralized registry for trust. The larger the ecosystem relying on such identifiers, the greater the risk of this single point of failure.
- **High cost** - Multiple redundant identifier systems add costs that could be avoided with an interoperable decentralized identifier standard.

Together, these challenges limit the use cases to which identifiers can be applied, thus slowing the propagation of new research and limiting the agility of research ecosystems.

The Opportunity and Solution

To increase the usability of identifiers we propose to describe a new ecosystem of Decentralized Resource Identifiers (DRIs). This ecosystem should be based on open standards for interoperability, capable of global scale, and highly adaptable to new and existing use cases. It should also reuse existing technologies whenever possible.

A DRI ecosystem can include tools, open source software, support, knowledge, training, examples, best practices, and governance structure examples. Organizations and their communities will be empowered to produce their own systems of identifiers and fully control and maintain them to best suit the group's needs. Rather than a prescribed, inflexible, centrally-controlled approach, the DRI

ecosystem can establish a set of open standards-based specifications describing how identifiers can be created, managed, resolved, and verified. This would give communities the freedom to create and maintain their own identifier systems which are interoperable and compatible with other systems. These new DRIs can be efficiently made backward compatible so they can be interoperable with existing databases and registries.

The following use cases help illustrate the opportunities:

Use case 1: An organization identifies the need in their community for global, permanent, resolvable, trusted identifiers for their resource. They develop a system of identifiers that best meets their needs including machine-actionable semantics, governance and trust.

Use case 2: Researchers can use multiple research identifiers to identify and cite equipment usage, grant funding, reusable code, publications etc. Through usage of the identifiers they can establish their reputation within the research ecosystem and monitor the impact of their work. Government institutions can link existing resources and measure the impact of financed research more easily because of the standardized usage of identifiers. Through this they can create better policy using more accurate information.

Use case 3: Machine Learning (ML) and Artificial Intelligence (AI) research has a potentially huge impact on the society as a whole beyond academic research or technical development. But this opportunity also comes with high risks of negative social cost such as: (1) biases in AI decision making, (2) privacy protection in collecting and handling large datasets in AI research, and (3) reproducibility challenges in AI research.

Decentralized Resource Identifiers can be designed with a decentralized registry to address some of the important issues to encourage and facilitate responsible AI research. Similar issues are also present in other scientific as well as social science research fields.

A DRI goes beyond what traditional identifiers can achieve and can help fundamentally improve research methodologies and toolsets to AI and other data science research. DRIs and verifiable credentials can support the tracking of research artifacts, including but not limited to collaborators, papers, datasets, data sources, test results; can provide verifiable transparency and accounting; can support biases verification; and can enforce privacy protection rules.

Use case 4: The Trust over IP (ToIP) community has a need for DRIs for the identification of all outputs including concepts, terms, mental models, examples, white papers, and other deliverables. Using such identifiers with appropriate syntax in 'raw' texts or markdown documents allows for tracking of usage and the automated creation of nicely rendered deliverables, that come with pop-ups of defined terms, automatically generated glossaries that explain the words used in the document, etc.

Decentralized Resource Identifier Ecosystem

An ecosystem of service, technologies and support for DRIs will enable global interoperable solutions allowing anybody to create, control and maintain their own persistent research identifiers at low costs. Identifiers could be applied to any digital content as well as any physical object which can be cryptographically linked with the identity of its manufacturer, provider and owner.

The ecosystem we will describe can consist of two parts: The infrastructure and the application of the infrastructure. This can include:

Guiding principles: The ecosystem should adopt principles that would support the community such as a commitment to current standards, interoperability, open-source development, and FAIR 2 data (Findable, Accessible, Interoperable, Reusable). A focus on ensuring compatibility with existing schemas would simplify conversions and ultimately identifier harmonization.

Governance examples: For any system of identifiers, a governance framework is needed to establish trust and overall language. The ecosystem can provide resources and examples of governance frameworks that can be adopted by any participating organizations looking to establish their own systems of identifiers.

Semantics: The ecosystem can supply examples, education, support, and semantic recommendations such as accessible and reusable schemas. Examples include Overlay Capture Architecture 3 or the Metadata 4 Machines workshops 4 organized by GoFAIR.

Training, documentation and promotion: All efforts in increasing usability of resource identifiers will require ecosystem support. These can include training sessions, resource documentation, and promotion and outreach at events such as the Canadian Science Policy Conference.

Technologies: The decentralized resource identifier ecosystem can leverage existing technologies developed in different communities including:

- W3C Decentralized Identifiers (DIDs) 5. The W3C DID Working Group, launched in September 2019, is nearing completion of the DID Core Specification for globally interoperable decentralized identifiers that are generated and verified cryptographically so they do not require centralized registries or service providers. The DID specification allows specific DID methods to be developed to support different decentralized verifiable data registry systems (blockchains, distributed ledger technologies, distributed file systems, peer-to-peer networks, etc.) Over 70 DID methods have been registered in the W3C DID Specification Registries 6, and several global-scaled DID networks have been implemented including the Sovrin network and the EU IDUnion network.
- KERI 7 (Key Event Receipt Infrastructure) can be used as a core technology to provide decentralized secure root-of-trust based on cryptographic self-certifying identifiers. It uses hash chained data structures called Key

Event Logs that enable ambient cryptographic verifiability. In other words, any log may be verified anywhere at any time by anybody. It has separable control over shared data which means each entity is truly self-sovereign over their identifiers. With KERI it is possible to create immutable, portable Decentralized Resource Identifiers which do not require centralized authority nor registry and can be used across all systems and use cases. To be able to resolve any identifier which is created with KERI there is a need for decentralized infrastructure. The resolution infrastructure is based on DHT (Distributed Hash Table) due to the properties of KERI which provides end-to-end verifiability we don't need to trust the location of the identifier's event log.

- ISCC 8 - Content Identifiers - ISCC identifiers are generated algorithmically from the content itself. Content files are processed to build the identifier. The ISCC does not have to be manually assigned, neither does it have to be carried around or embedded within the content. The content itself is the source and authority of the ISCC Code. The ISCC Code is a unique, hierarchically structured, composite identifier. It is built from a generic and balanced mix of content-derived, locality-sensitive and similarity-preserving hashes generated from metadata and content.

- The Verifiable Credentials trust triangle 9: This architecture establishes the three core roles for transitive digital trust: issuers, holders and verifiers of digital credentials. Digitally-signed credentials of various kinds are used today with various types of identifiers to link data objects—for example linking an employee ID to a research paper published by the university. Unfortunately some of these identifiers (e.g. employee ID) lose their meaning as soon as they leave the domain in which they were created. Using Decentralized Resource Identifiers, we can solve that problem by having uniquely global identifiers which are resolvable outside of the domain where they were created. This improves interoperability of linked data and enables the trust triangle to be portable and transitive.

- The Ceramic Protocol 10: This protocol provides a decentralized document storage with versioning and multiple ownership. Each document has a DID permalink and at least one owner DID. The network builds up a graph of versions of the document and uses cryptographic signatures and anchoring on a blockchain to track and resolve official versions. The protocol uses its own DID method labelled 3ID to reference accounts and to connect them across blockchains. A Ceramic document can have links to other documents that are referred to as tiles. The tiles allow the document to provide relevant information about the document, how it can be used, any services associated with it, versioning and the owners. The tiles allow researchers to link together their paper, references, any code or services they used along with their data sets. The documents are stored in a distributed system so that they are always available and the link is permanent. The protocol is public and permissionless, censorship resistant

and resilient.

- Blockchain/Distributed Ledger Technologies (DLT) 11: Blockchain, and more broadly DLT, is an emerging technology that provides a decentralized 'write only' ledger to record data events and identifiers. In this way, blockchain provides a method of ensuring the provenance of data via temporality (time stamps), and trustworthiness as new information can only be appended, not overwritten, to form a resilient and immutable record.

A research identifier ecosystem taking advantage of new technologies such as decentralization, credentials, DIDs, hashes and cryptographic verification can achieve benefits such as interoperability, security, and scalability compared to current implementations.