The opportunities of Decentralized Resource Identifiers in the research landscape

Carly Huitema (Manager, WCMR, University of Waterloo and ToIP member – contact carly.huitema@uwaterloo.ca), Robert Mitwicki (Head of the Technology Council at Human Colossus Foundation), Dave McKay (Technical Lead, Cybersecurity Research Lab at Ted Rogers School of Management at Ryerson University, Co-Chair Innovation Experts Committee at DIACC), Wenjing Chu (Senior Director of Open Source and Research, Futurewei Technologies, Steering Committee member of the ToIP Foundation), Dian Ross (PhD student, Blockchain@UBC)

Current issue

Identifiers of research objects that are persistent, unique, and globally resolvable are increasing in importance[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, but creation of each additional identifier system must begin *de novo*.

Identifiers are very 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 to the research ecosystem will be useful for every stakeholder of the digital research infrastructure initiative of NDRIO.

Existing identifier systems are useful but come with a set of challenges that make them difficult to implement for new use cases including:

> **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** leads to using multiple identifiers for the same object just to ensure compatibility between the systems. This redundancy introduces additional complexity and makes it harder to maintain.
>
> **Fragmentation** - Different types of identifier systems each have their own specific implementations, increasing the costs of development and maintenance and making integration between systems much harder.
>
> **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 - A centralized identifier system centralizes costs as well. Multiple redundant identifier systems also 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, slowing the propagation of new research and limiting the agility of research ecosystems.

The opportunity and solution

To increase the usability of identifiers we need to build a new ecosystem of decentralized identifiers. This ecosystem should be based on open standards for interoperability, capable of global scale, and highly adaptable to existing and new use cases. It should also reuse existing technologies whenever possible.

As the national organization for digital resources in research, NDRIO is the best organization in Canada to establish such an ecosystem of expertise, tools, and support for research communities to create and use the persistent resource identifiers they need. NDRIO can become a first mover in ***creating an ecosystem of Decentralized Resource Identifiers for research that are easier to create and maintain, lower cost, more flexible, and eliminate the danger of single points of failure.***

NDRIO can then empower and support organizations and their communities 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, NDRIO can establish a set of open standards-based specifications within a decentralized ecosystem 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 Decentralized Resource Identifiers can be quite efficiently made backward compatible so they can be interoperable with existing databases and registries.

As the Decentralized Resource Identifier ecosystem champion, NDRIO can produce on their own or in partnerships: tools, open source software, support, knowledge, training, examples, best practices, and governance structure examples. This will allow groups who recognize the need for identifiers in their community to rapidly, cheaply, competently, sustainably, and with lower risk, execute their own implementations while encouraging standards, best-practices, reuse, and interoperability.

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. Working with NDRIO they develop a system of identifiers that best meets their needs including the machine-actionable semantics, governance and trust. They choose to work with NDRIO to support any necessary infrastructure they may need to run the system, perhaps connecting them internationally to a network of nodes all supporting these Decentralized Resource Identifiers.

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

Use case 3: Responsible Artificial Intelligence (AI). Machine Learning (ML) and AI research has a potentially huge impact on the society as a whole beyond academic research or technical development. But this huge 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, (3) reproducibility crisis 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 Decentralized Resource Identifier goes beyond what traditional identifiers can achieve and can help fundamentally improve research methodologies and toolsets to AI and other data science research. Decentralized Resource Identifiers and verifiable credentials can support 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 enforce privacy protection rules.

NDRIO can support the ecosystem further with competitive grants for researchers to establish their own identifier systems using the NDRIO ecosystem. NDRIO can also incorporate the ability to cite identifiers (**all** identifiers) into **all** NDRIO supported products to ease the burden of accurate citations and encourage identifier uptake. NDRIO can be a global leader in implementing an ecosystem of Decentralized Resource Identifiers to the benefit of all research enabling the research community to organize independently, build critical mass, and ultimately execute at scale.

NDRIO's Role

NDRIO can create and nurture an ecosystem of service and support for Decentralized Resource Identifiers. With these identifiers it would be possible to provide global interoperable solutions which would allow anybody to create, control and maintain their own persistent 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 will consist of two parts: The infrastructure and the application of the infrastructure. NDRIO can create, share, and operate the decentralized infrastructure which will support the use and integration of resource identifiers that operate in an interoperable and global way. NDRIO can also support applications using this infrastructure through the development of parts of the ecosystem including examples, training, documentation, funding, etc.

Components of the ecosystem 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. As leader of the ecosystem, NDRIO 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: NDRIO 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: NDRIO can leverage existing technologies developed in different communities to bootstrap the Decentralized Resource Identifier ecosystem 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 enables 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.

Recommendations

1. Beginning with a Proof of Concept and using some or all of the technologies described here, NDRIO should create an identifier ecosystem supporting open source, open standards, decentralization, and interoperability.
    a. Initial use cases to consider of varying scope and impact include: The Canadian MetaMicrobiome Library[12], iGEMs Registry of Standard Biological Parts[13], and grant identifiers for Tri-Council grants.
2. NDRIO should incorporate all digital resource identifiers into their workflow and products.

[1] https://datascience.codata.org/articles/10.5334/dsj-2020-046/

[2] https://www.go-fair.org/fair-principles/

[3] https://oca.colossi.network/

[4] https://www.go-fair.org/resources/go-fair-workshop-series/metadata-for-machines-workshops/

[5] https://www.w3.org/TR/did-core/

[6] https://www.w3.org/TR/did-spec-registries/

[7] https://keri.one

[8] https://iscc.codes/

[9] https://trustoverip.org/

[10] https://ceramic.network/

[11] https://blockchain.ieee.org/standards

[12] http://www.cm2bl.org/

[13] http://parts.igem.org/Catalog