IJDC | *General Article*

A System for Distributed Minting and Management of Persistent Identifiers

Łukasz Bolikowski
ICM, University of Warsaw

Aleksander Nowiński
ICM, University of Warsaw

Wojtek Sylwestrzak
ICM, University of Warsaw

Abstract

Minting persistent identifiers and managing their metadata is typically governed by a single organization. Such a single point of failure poses a risk to longevity and long-term preservation of identifiers. In this paper we address the risk by proposing a radically different approach, in which minting and management of persistent identifiers is distributed, and the integrity of the distributed system is guaranteed by public-key cryptography. We describe the general architecture of the system, analyse its robustness and discuss potential deployment scenarios.

Received 13 February 2015 | *Accepted* 13 February 2015

Correspondence should be addressed to Łukasz Bolikowski, ICM UW, Prosta 69, 00-838 Warszawa, Poland. Email: l.bolikowski@icm.edu.pl

An earlier version of this paper was presented at the 10th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

A proliferation of research artifacts in contemporary scholarly communication, in particular the publication of datasets and source code mandated by the principles of open science, has emphasized the need for minting and management of persistent identifiers. Traditional, centralized approaches guarantee persistence via trusted organizations with long-term commitment to identifier management. However, this model has inherent risks, as identifier persistence is dependent on the viability, solvency, and good will of the managing organizations, and their ability to survive political, economic and military turmoils. Even in favourable external conditions, the above risks have already materialized several times in the short history of the Internet (cf. shutting down of MyOpenID). Therefore, even the most technologically robust identifiers tend to have a single point of failure: their managing organization. A handful of initiatives were recently launched to address these challenges including w3id, which relies on a social contract between several managing organizations.

Goal

In this paper we propose a radically different, decentralized scheme for minting and management of persistent identifiers. We drew inspiration from two popular, distributed systems based on cryptography: Git for source code management and Bitcoin for online payments. Our goal is to design, implement and deploy a system with the following properties:

- Anyone can mint a persistent identifier and associate with it a URL to content and a list of keys authorized to manage the identifier;
- For a given persistent identifier, any authorized key owner (and no one else) can alter the URL and the list of authorized keys;
- Anyone can download the complete set of persistent identifiers with complete revision history and verify its integrity;
- No proper subset of participating people and organizations is capable of shutting down the system.

The focus of this work is, therefore, specifically on maintaining long-term resolvability of persistent identifiers. The system should be agnostic with respect to referent type (data sets, source codes, documents, people) and content delivery technology (HTTP, BitTorrent, Tor/Onion). Thus, the system adopts the PID paradigm (Van de Sompel et al., 2014) and will be ready for any future referent types and technologies.

Method

To achieve our goal, we propose Peer-Minted Persistent Identifier (PMPI) system that makes extensive use of public-key cryptography and peer-to-peer network communication. For the sake of exposition, we will present here a slightly simplified description of the system.

 doi:10.2218/ijdc.v10i1.368

The key notion in the system is an **operation** (such as a mint or update) on an identifier. Each operation contains the following essential fields:

- An operation identifier (a hash of all the other fields),
- A persistent identifier (a UUID),
- A URL to content,
- A list of public keys of entities authorized to manage the identifier,
- An identifier of the preceding operation,
- The digital signature of the authorized person or organization executing the operation,
- A proof of work (cryptographic nonce).

Each operation (save for the root) has a predecessor, and the majority of the operations in the system form a **chain** (see Figure 1), while occasional "orphaned" operations will exist outside the main chain due to failed concurrent modifications. An operation can be added to the chain if it is valid, which means in particular that the digital signature is correct and the private key used for signing is authorized to manage the identifier (i.e., the most recent operation in the chain related to the persistent identifier features the corresponding public key), etc.

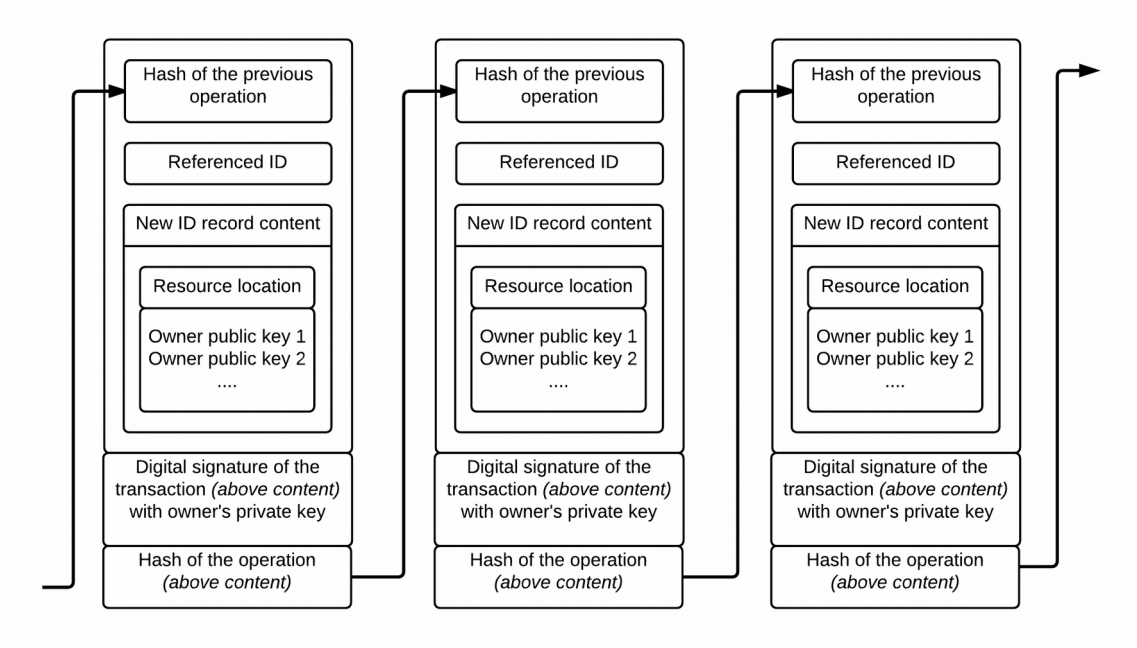

Figure 1. A chain of operations on persistent identifiers.

The proof of work is a solution to a cryptographic puzzle. The executor of the operation has to invest a moderate amount of CPU time to solve the puzzle and come up with a valid operation. This serves several important purposes: it helps to sequence operations performed concurrently, and reduces incentives for spamming and flooding attacks. Entities that minted a significant number of identifiers (and therefore invested significant computational power) have a higher "reputation" and are given easier

puzzles, thus facilitating bulk operations by registrars and large libraries. More precisely, a trusted user can put several operations in a single block added to the chain, while producing a single proof-of-work. The more trusted a user, the more operations they may put in a block.

Certain decisions in the system should be taken collectively, rather than by individual users. For example, when a given key is consistently used for spamming or other malicious activity, the community may choose to blacklist the key. Or, when the only user authorised to control a certain set of identifiers loses their private key, the community may collectively allow them to reclaim control over the orphaned identifiers. In both of these examples a certain level of consensus should be required in order to perform a globally-beneficial action. Therefore, in order to facilitate collective decision-making, another type of operations is added to the system (these operations are chained and signed just like any other operation): votes supporting certain well-defined types of notions.

PMPI incorporates a system of roles. Users authorised to act on a given identifier may be given different access rights, for example an organization providing metadata curation services may be authorized to change the checksum of the referent, but cannot modify the list of authorised keys.

PMPI has a peer-to-peer architecture (see Figure 2). Each node in the system keeps the complete set of operations. A new operation (such as minting an identifier) is created by linking to the most recent operation in the main chain. Immediately after a proof of work is found, the operation is broadcast to the network. Each receiving node verifies the operation and stores it. This way, the entire system can be restored from a single node, and any kind of tampering with any operation in the chain can be easily discovered (e.g. digital signatures will not match, or the operation identifier will not match the one stored in its successor).

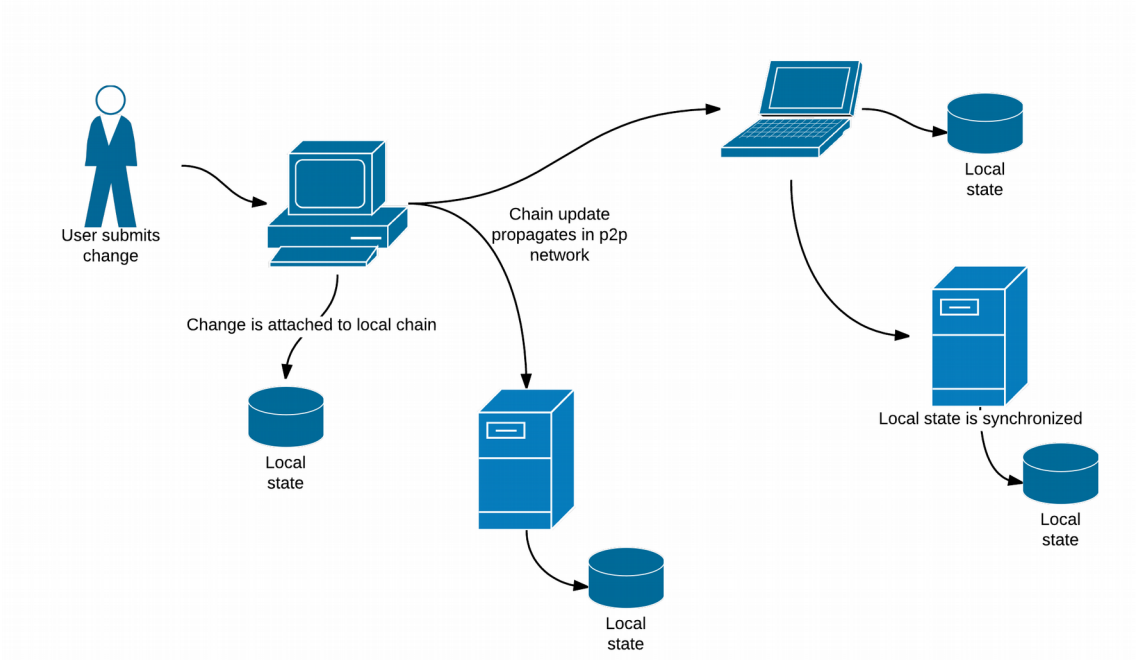

Figure 2. Information flow in P2P network.

 doi:10.2218/ijdc.v10i1.368

Properties

The proposed system follows the LOCKSS (Lots of Copies Keep Stuff Safe) way of providing long-term preservation. Each node in the system contains its entire state, which incurs only modest space requirements (back-of-the-envelope estimation: 500 million identifiers with two operations per identifier on average and 400 bytes per operation means 400 GB of storage needed).

PMPI is transparent and secure: anyone can read the entire revision history of any persistent identifier, and anyone can verify the integrity of the entire chain of operations, as verifying proofs of work and digital signatures is cheap.

Business Models

For-profit and non-profit organizations may offer a range of services based on PMPI, from minting identifiers (thus monetizing their good reputation in the system), to maintaining the resolvability of identifiers (as part of their data stewardship services), to offering backup access (providing insurance against the loss of a private key), to providing metadata curation services on top of minting persistent identifiers.

Resilience to Attacks

The system is designed with security in mind, including the ability to withstand foreseeable types of malicious attacks is a priority.

One of the foreseeable vectors of attack is a denial of service caused by flooding the system with operations (e.g. minting new identifiers or acting on existing ones). As each operation is stored by all the nodes in the system, the machines would eventually run out of disk space. Such an attack is fortunately infeasible thanks to the proof-of-work: it takes too much more time for the adversary to mint a new identifier.

Even if an entire botnet were employed to disrupt the system, it could only succeed by first generating a single high reputation key and thus being able to execute bulk operations. This key, however, would be blacklisted (nodes would stop accepting operations executed using the key) and the computational power of the botnet would be wasted. Coordination of the blacklisting would have to be done outside the protocol (e.g. via a mailing list of PMPI operators), but on the other hand the attack itself seems to be extremely unlikely, given the amount of investment necessary and the lack of benefit.

Access Control

As in any system based on public-key cryptography, losing one's private keys can have very negative consequences. Fortunately, there are mechanisms in PMPI that mitigate the risk of losing access to one's identifiers. First of all, one can authorize another key owner (one's organization, or some trusted third-party organization) to co-manage a given set of identifiers. The protocol also gives the community access to abandoned identifiers.

Interoperability

PMPI is not intended to eliminate the existing minting organizations, as many of them provide additional services that are not provided by our system (for instance, handling and curating metadata). Instead, we seek to provide an interoperability mechanism through which the end users will be able to migrate from one provider to another, while retaining the same persistent identifier. Those who choose to do so, will be able to mint and manage identifiers on their own, others will be able to delegate specific concerns to the service providers of their choice.

Related Work

No solution known to us shares the goals and characteristics of PMPI. Existing persistent identifier solutions obviously share the goals of PMPI, but are not distributed, and therefore long-term preservation of identifiers is dependent on the viability and goodwill of managing organizations. However, there are several solutions with different goals, but similar approaches (distributed, based on public-key cryptography and Merkle trees). A careful reader might ask: "Why not take solution X and use it for our benefit?" Let us answer this question for several most likely values of X.

Namecoin

Namecoin[1] is a direct fork of Bitcoin[2], with a goal of resisting online censorship. It is a "decentralized open source information registration and transfer system". However, while Namecoin focuses on fighting censorship, PMPI puts more emphasis on long-term preservation aspects: recoverability of identifiers, access control and delegation. There are further differences between PMPI and Namecoin that are not specific to the latter and apply to all Bitcoin-based solutions.

Permacoin

Permacoin (Miller et al., 2014) aims at repurposing Bitcoin's proof-of-work for data preservation. It is geared towards larger pieces of information that are to be stored in a distributed manner. So-called proofs-of-retrievability are used to check whether users really do store fragments of the public data set. PMPI, in contrast, deals with many small and mutable pieces of information and has more fine-grained access control mechanisms.

Bitcoin

Bitcoin (Nakamoto, 2008; Decker and Wattenhofer, 2013) is the most popular digital cryptocurrency, and as of early 2015 its market capitalization is in the order of billions of U.S. dollars. Bitcoin was one of the inspirations for PMPI, but there are several important differences between the two. In contrast to Bitcoin-based systems, PMPI does support **bulk minting operations**, by loosening the requirement of presenting proofs-of-work in the case of reputable key owners. More broadly, the primary purpose of a proof-of-work in Bitcoin-based systems is to generate a digital currency (which is convertible to U.S. dollars), while PMPI decouples financial aspects of minting from

1 Namecoin: http://namecoin.info/
2 Bitcoin: https://bitcoin.org/en/

technical ones. In other words, PMPI guarantees security and integrity of minting, but does not impose any particular business model.

Furthermore, while Bitcoin and its forks support "compressing" blocks of old operations, in PMPI the variable size of blocks serves one additional purpose: it rewards reputable users, making it easier for them to perform bulk operations. Moving on, neither Bitcoin nor its forks support collective decision-making (blacklisting known offenders, recovering from loss of private keys). Finally, PMPI facilitates delegation of concerns (different key owners may have different rights with respect to the same identifier).

Conclusions

We propose a system for the minting and management of persistent identifiers, in which long-term preservation of information is no longer dependent on any single organization, but instead on the existence of many publicly available copies. By default, identifier owners have complete administrative control by means of public-key cryptography, but can delegate that control to for-profit or non-profit organizations. Therefore, the proposed system increases transparency and public access to information about identifiers, while retaining viable business models for registrars.

Current Status and Next Steps

PMPI is a well-studied idea for a persistent identifier system, with a lot of effort put into guaranteeing its security, scalability and robustness. We plan to identify and gather stakeholders interested in developing a concrete protocol for PMPI. Our primary focus will be on scholarly communication and the various types of entities commonly referenced there, such as documents, data sets, source codes, people and organizations. Together with a protocol, we plan to develop a reference implementation of P2P software for the distributed management of persistent identifiers.

References

Decker, C., & Wattenhofer, R. (2013). Information propagation in the Bitcoin network. In *2013 IEEE 13th International Conference on Peer-to-Peer Computing*. doi:10.1109/P2P.2013.6688704

Miller, A., Juels, A., Shi, E., Parno, B., & Katz, J. (2014). Permacoin: Repurposing Bitcoin work for data preservation. In *2014 IEEE Symposium on Security and Privacy* (pp. 475–490). doi:10.1109/SP.2014.37

Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system.* Retrieved from https://bitcoin.org/bitcoin.pdf

Van de Sompel, H., Sanderson, R., Shankar, H., & Klein, M. (2014). Persistent identifiers for scholarly assets and the web: The need for an unambiguous mapping. *International Journal of Digital Curation, 9*(1), 331–342. doi:10.2218/ijdc.v9i1.320