F O U N D A T I O N S O F C O M P U T I N G A N D D E C I S I O N S C I E N C E S
Vol. 44 (2019) No. 4
ISSN 0867-6356
DOI: 10.2478/fcds-2019-0018 e-ISSN 2300-3405

sciendo

A blockchain based Trusted Persistent Identifier system for Big Data in Science

Emanuele Bellini *

Abstract. A stable reference of Internet resources is crucial not only to identify a resource in a trustworthy and certified way but also to guarantee continuous access to it over time. The current practice in scientific publication as the use of a Persistent Identifier (PID) like a DOI or Handle, is becoming attractive also for the datasets. In fact, in the era of Big Data, the aspects of replicability and verification of the scientific result are paramount. In this paper we verify the functional feasibility of permissioned blockchain technology as a tool to implement a Trustworthy Persistent Identifier (T-PID) system for datasets in the scientific domain.

Keywords:
Persistent Identifiers, Big Data, Blockchain, Hyperledger Fabric, Trust

1. Introduction

Today, Open Science and reproducible research have become a crucial goal across research communities and funding bodies [18], [6], [32]. According to the Joint Declaration of Data Citation Principles (JDDCP) [27], the first of the core principles on purpose, function, and attributes of data citations, is that data should be considered legitimate, citable products of research [3]. The JDDCP requires that data become a first-class research object that is archived in persistent stores and is cited similarly as a publication, in all cases where: (1) findings or claims are based on the author's primary research data; or (2) data from other sources is input to the author's analysis. The 3rd JDDCP Principle requires also that wherever research findings are based upon data, that data be cited, while the 4th Principle requires that cited, archived data receive a globally unique, machine-resolvable Persistent Identifier (PID) that should appear in the reference of the citing article. This is intended not only to

*Center of Cyber-Physical System, Khalifa University, Abu Dhabi, UAE, emanuele.bellini@ieee.org, ORCID:0000-0002-7878-8710

help humans locate data but to facilitate the development of next-generation mashup tools in an ecosystem based on software agents and searchable research data indexes such as DataMed and OmicsDI[51]. The importance of using a PID emerged in [13], where the resulting low quality of descriptive metadata in academic Open Archives limits the ability to retrieve information. In turn, a survey conducted to profile the importance of the Dublin Core fields for content retrieval, revealed that DC:Identifier field was as important as fields that represent the pillars of information retrieval in the Digital Libraries like DC:Title and DC:Author [15]). Moreover, it is well known that the structural instability of URLs and related resources because of relocation or updating, is one of the main issues that affect the information retrieval possibility on Internet. In fact, we have many documented issues related to the URL usage as an identifier (instead of a simple locator) such as:

- the typical loss of reference to digital object that leads to the famous HTTP 404 error when the object is moved on another server;
- the loss of confidence about dataset authority and provenance;
- the difficulties to recognize if the dataset has been manipulated (e.g. by versioning).

Today, a number of initiatives, standards, and technologies are available as PURL, DOI, NBN, Handle System, Cool URI, Linked Open Data, ARK. However, not all of them are suitable for being used with Big Data. According to the review conducted in [29], the solution considered mature for data identification are the following: DOI, URN, ARK, Handle, PURL and Accession number. Even if the DOI can be considered the standard de-facto for scientific products (included data), there are several well-established identification mechanisms in some communities that do not seem to be about to be discharged. For instance in biomedical data, which is composed of over 600 autonomous repositories independently funded, DOI is not adopted[45]. Moreover, the level of service in terms of granularity, namespace management, identification policies, etc. could be very specific so that general purpose PID system may have difficulty in meeting these requirements. This is evident, in cultural heritage domain where DOI adoption could encounter difficulties m in being adopted because of its for-profit vocation and lack of policy for objects preserved in a long run.

On the contrary, services like the IETF RFC-3188 based Italian National Bibliography Number (NBN:IT) managed by the Central National Library of Florence, is specifically designed to identify only preserved digital objects. This policy makes the PID not very useful for intensive scientific dissemination.

Finally, it is necessary to consider that a PID Registration Agency (RA) is not aware if copies of the same dataset have received PIDs from other systems [14]. This scenario contributes to making difficult the recognition of the right dataset.

It is clear that we have to consider the PID as a socio-technical system that involves several actors, technologies, business processes and scopes. On the other hand, every PID system on Internet should be able to redirect users to reliable resources while guaranteeing the truthfulness of assertions (implicit or explicit) on its localization, type of content, provenance, authority, etc.. Moreover, a PID system has to

assure a long term commitment to maintain these assertions and related services (e.g. resolution). In fact, the citability issues is not imitated to the technological aspects. Up to now, several PID systems and initiatives have failed to withstand the test of time, sliding into paralysis and a "zombie" stage as described in [7], [34].In this case, identifiers may continue to exist but the resolution service is not active anymore. According to [30] there are several factors contributing to the failure that include:

- complexity of the PID systems;
- lack of long-term financial support from hosting organisations;
- narrow target community; and
- reliance on a single governance authority or hosting organisation.

Such considerations can be classified according the Bricks of Trust model (BoT) for PIDs, as proposed in [9]. In particular, the complexity of the system concerns the Technology (T) and Organizational (O) Bricks in BoT. The systems should be designed to be scalable, resilient, flexible, extendable and maintainable at technological level (T). Moreover, a PID system should have clear governance that needs to be reflected in the technological architecture (O). The lack of Financial support is related to a failure in the Financial Sustainability (F) brick that is referred to the loose of the economic support for many reasons such as organizational objectives and missions changing, market crisis, and so forth. However, the most critical situation for the PID survival occurs when the service provider withdraws its commitment or when the community served withdraws its mandate. Even if for the former may exist a recovery plan in case another organization replace the previous one guaranteeing the business continuity of PID service, the latter is a dead end road.

This paper aims at discussing these aspects under the Big Data perspective and explores the possibility of using permissioned blockchain as a tool to implement a trust PID system (T-PID)for Big Data in Science. One of the first attempt to consider blockchain as a suitable technology to implement PID system has been explored in [17]. However, even if there has been growing interest in exploring the actual capacity of this emerging technology to implement a PID system [1], the main functions supplied by a PID system are not seen as sequential in nature yet[22].

Hence, in the present work we want to make such benefit evident highlighting the current challenges in Big Data persistent identification and, how these challenges, can be positively addressed with a permissioned blockchain approach. In this respect the article is structured as follow: in section 2 a summary of the related work is provided; in section 3 the most relevant challenges related the Big Data persistent identification in Science are discussed; section 4 is dedicated to design and explore the use of a permissioned blockchain to implement a trusted Persistent Identifiers (T-PID) for Big Data in Science; conclusions are provided in section 5.

[1] https://orcid.org/blog/2018/03/07/why-we-need-explore-blockchain-technology-connect-researchers-and-research

2. Related work and initiatives

In this section are presented the main relevant PID systems and emerging approaches. In [28] has been assessed and compared nine technologies and systems for assigning persistent identifiers for their applicability to earth science data (ARK, DOI, XRI, Handle, LSID, OID, PURL, URI[2], URN[3], URL, and UUID[4]) depending on their technical, user and archival values.

The Document Object Identifier system (DOI) [42] is a business-oriented solution widely adopted by the publishing industry, which provides administrative tools and a Digital Right Management System (DRM). The DOI System is currently being standardized through the ISO/DIS 26324, Information and documentation Digital Object Identifier System and its syntax it has been standardized also as ANSI/NISO Z39.84-2005, that prescribes the form and sequence of characters comprising any DOI name. A DOI name is permanently assigned to an object, to provide a persistent link to current information about that object, including where it, or information about it, can be found. The principal focus of assignment is to content related entities that include data sets. Archival Resource Key (ARK [37]) is an URL-based persistent identification standard, which provides peculiar functionalities that are not featured by the other PI schemata, e.g., the capability of separating the univocity of the identifier assigned to a resource, from the potentially multiple addresses that may act as a proxy to the final resource. The Handle System [5][44] is a technology specification for assigning, managing, and resolving persistent identifiers for digital objects and other resources on the Internet. The protocols specified to enable a distributed computer system to store identifiers (names, or handles) of digital resources and resolve those handles into the information necessary to locate, access, and otherwise make use of the resources. That information can be changed as needed to reflect the current state and/or location of the identified resource without changing the handle. The Persistent URL (PURL) [6] is simply a redirect-table of URLs and it is up to the system-manager to implement policies for authenticity, rights, trustworthiness, while the Library of Congress Control Number (LCCN) [7] is a persistent identifier system with an associated permanent URL service (the LCCN permanent service), which is similar to PURL but with a reliable policy regarding identifier trustworthiness and stability. Among the URN based PIDs, the most famous one is the National Bibliography Number-NBN (IETF RFC 3188) developed in the context of the National Libraries in Europe. In particular, the Italian project [8] [9]), started in 2007, assigned the NBN:IT identifier, to the digital resources deposited for the long term preservation. In this respect, the namespace NBN:IT assumes an iconic value that informs the user about the fact that the resource identified is not only retrievable online but that there is a copy preserved. Even if there is not a specific technical limitation, the

[2]https://tools.ietf.org/html/rfc3986
[3]https://tools.ietf.org/html/rfc8141
[4]https://tools.ietf.org/html/rfc4122
[5]Handle System website, http://www.handle.net/
[6]Persistent URL http://purl.oclc.org
[7]Library of Congress Control Number http://www.loc.gov/marc/lccn.html
[8]see http://www.depositolegale.it/national-bibliography-number/

service does not consider datasets a type of object eligible for being preserved by the National Library.

Another interesting approach is provided in [50]. Two technological implementations to support persistence have been confronted: BitTorrent and name Data Network. BitTorrent, a well-established location independent access technology. It works on top of today's location-based networks with Transmission Control Protocol (TCP) and User Datagram Protocol (UDP). In contrast to existing location-based data repositories that are subject of PID target resolution, BitTorrent technology uses a peer-to-peer approach supporting parallel downloads. With its latest features of Distributed Hash Table (DHT) and Peer Exchange (PEX), BitTorrent does not require central infrastructure to discover other network peers and localize files [38], [39]. BitTorrent uses infohashes that are computed as SHA-1 check-sums on the content of the file. Every peer that possesses the infohash can download the data set from the BitTorrent swarm that consists of the peers offering the data set for download. The swarm arrangement and the overlay network for the specific file are computed for every download [50]. The second approach is represented by Named Data Networking (NDN) is a current research topic of location-independent data access using information-centric principles. NDN is also featured in the location-independent PID approach presented in this paper to support a next-generation Internet technology. In NDN, data sets are enumerated through Data Names that form a hierarchical namespace [35]. This overview shows that it is not viable to impose a unique PID technology and that the success of the solution is related to the credibility of the institution that promotes it. Moreover, the granularity of the objects that the persistent identifiers need to be assigned to is widely different in each user application sector. At EU level PIDs issues have been addressed by several projects as DigitalPreservationEurope, APARSEN [9], ODIN [10], THOR [11] and the current FREYA [12]. Also, national and international initiatives emerged. Among them is valuable to mention PersID [13], composed by European National Libraries and foundations to create an interoperability framework among PID-S, ePIC [14] and the current studies and working group established at DRA, CESSDA, and EUDAT.

Another emerging initiative is represented by the Decentralised Identifier (DID) supported by W3C [15].

DID that like a URL, can be resolved or dereferenced to a standard resource describing the entity, but unlike a URL, the DID Document typically contains cryptographic material that enables authentication of an entity associated with the DID. DID is also intended to be persistent, the specifics recommend that "DID method specifications only produce DIDs and DID methods bound to strong, stable ledgers or networks capable of making the highest level of commitment to persistence of the DID and DID method over time", without any specific technological and organiza-

[9] http://www.alliancepermanentaccess.org
[10] https://odin-project.eu
[11] https://project-thor.eu/
[12] https://www.project-freya.eu
[13] http://www.persid.org/
[14] (https://www.pidconsortium.eu/)
[15] https://w3c-ccg.github.io/did-spec/

tional reference for its implementation. Among the emerging technologies a special consideration needs to be dedicated to the IPFS [16]. The IPFS technology has been explored also to directly implement a persistent identifier system as reported [48]. In fact, one of the core feature of IPFS is that the links don't change because the system is based on the Content Identifiers (CID). The link does not indicate the location of the content, but it forms a kind of address based on the content itself. Any differences in the content will produce a new CID and the same content replicated along the IPFS nodes will have the same CID. At first glance. it seems that most of the features of a PID system can be implemented with IPFS. However the IPFS is a closed world, so that it is necessary to migrate the data management system towards this technology in order to benefit of a persistent identification of the datasets. Thus, despite the technology is promising, the need of a relevant technological shiftment in data management might represent a barrier difficult to overcome.

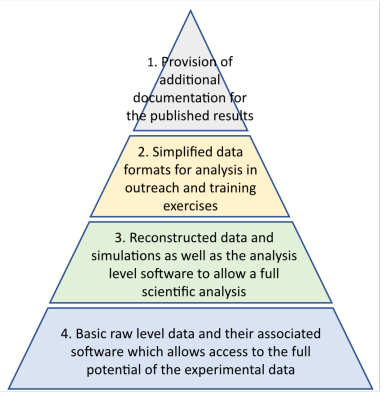

Figure 1: Data citability pyramid

A practical example of Data citation is provided by DATACite DOI Agency. DataCite DOIs are intended as citation identifiers for datasets. They inherit the same level of service of the DOIs but provide a standardized metadata schema, including links to associated publications, contributors and founders. The goal of DataCite is to enable data integration/mashup among heterogeneous datasets from diversified domain repositories and make data discoverable, accessible, and usable through a machine-readable and actionable standard data infrastructure. Thus, they focus on the data in the top section of the pyramid (*see* Fig. 1). Despite it is technically viable the use of DataCite DOIs for data belonging to the lower levels of the pyramid, other identifiers can be more appropriate because of a more focused purpose and level of service (e.g. Dynamic Data citation). In this respect, the current paper aims at exploring a blockchain based solution to cover the bottom part of the pyramid with a trustworthy PID system.

[16] https://ipfs.io/

3. Big Data persistent identification challenges in Science

Research data enable researchers to verify results and pursue new research questions. However, it might be critical to determine precisely which data were used to obtain a specific outcome and to be able to access the exact dataset. This is particularly true in the Big Data scenario, where algorithms process streams of data managed with unstructured data lakes. In this respect, PID systems are essential for data citation since metadata attributes are not able to unambiguously identify a particular item and cannot be used for a reliable location, retrieval or verification of research results. The ability to precisely reference datasets in a long term view is a key goal of data citation. This is relatively straightforward where data are static or fixed and their complexity limited. Challenges arise, however, when the data to be referenced are dynamic, complex and at large scale. In these cases, the simple adaption of a PID (as DOI or Handle for instance) could be not enough. With ever-growing data volumes, their processing into new datasets, and the need to maintain control over their quality and reliability [24], the use of unique and persistent identifier systems is necessary for their ability to version dynamic datasets and keep track of the transformations. As with publications, the PID type often depends on the exact data type and disciplinary background, and there is a great variety of established PID types for data. User communities in different disciplines need to decide on how to deploy PIDs going forward. As an example, in longitudinal studies, how to manage the versioning of the resulting datasets is a topic of debate. The issue about if a new PID is required for each version of a dataset or, in turn, related versions of a dataset should be archived with the same PID, cannot be easily solved. In fact, solutions may be different for social sciences respect to life sciences.

In geological and biological sciences, physical samples of biological or geological origin can be stored, processed and analyzed by various researchers with different scopes. In this way, the scientific outcome of a collected sample can increase. Furthermore, identification of the physical samples enhances the reproducibility of research allowing for quality control and tests of data from previous analyses. For these purposes, linking physical samples with publications and related data is crucial. Physical Samples from the natural environment may be identified with an ID in the form of the International Geo Sample Number (IGSN). The IGSN is a unique alphanumeric code, which is assigned through the registration of physical samples. IGSN was developed by the System for Earth Sample Registration (SESAR), which also issue IGSNs along with various other Allocating Agents following the same IGSN rules and regulations. IGSNs can be assigned to samples from a broad range of origin: rock, mineral, and fossil specimens, macro- and micro-biological samples and more. IGSNs are Handles that can be resolved to landing pages describing the sample. Elsevier and Copernicus earth science journals have recently implemented the use of IGSNs. Implementation of IGSNs is also recommended by the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS). PANGAEA has also long been publishing data and metadata that include IGSNs, thereby allowing for the data to be traced back to specific physical samples. The cross-linking between PIDs for physical samples and data is not an easy task. Even though IGSNs uses Handle System, there is no guar-

antee of a digital registration of the samples with the appropriate metadata assigned. Further difficulties revolve around PIDs for physical samples that follow sample ID systems other than IGSN, which may be specific to scientific disciplines, countries or institution. Currently, IGSNs have not been linked to DataCite and Schema.org metadata, but various initiatives are working towards expanding the implementation of IGSNs in international data management.

Big data workflows often require the assembly and exchange of complex, multi-element datasets. For example, in bio-medical applications, the input to an analytic pipeline can be a dataset consisting thousands of images and genome sequences assembled from diverse repositories, requiring a description of the contents of the dataset in a concise and unambiguous form. Typical approaches to creating datasets for big data workflows assume that all data reside in a single location, requiring costly data marshaling and permitting errors of omission and commission because dataset members are not explicitly specified[25]. In resilience and decision science, evidence driven decision making based on Big Data is emerging as the new trend [16]. Critical decisions are taken on the base of a huge amount of data collected from heterogeneous sources (peoples, sensors, communication networks, and so on). This poses relevant issues such as the possibility for an ex-post forensic investigation on the correctness of the decisions taken. In this respect, decision support systems should guarantee both identification and immutability of the datasets consumed for their analysis. Even Physics experiences the challenge of large experiment reproducibility also because of the difficult to make available online datasets with very large dimensions (Tera or Petabyte). The experiments at the Large Hadron Collider (LHC) are prominent examples. It is this uniqueness that makes the experimental data valuable for preservation and reuse with other measurements for comparison, confirmation or inspiration [26]. An example o such kind of Big Data, according to A Large Ion Collider Experiment (ALICE) experiment, are:

- Raw data embedding the signals delivered by the detectors along with the associated status data containing various information on the running conditions;
- Monte-Carlo data, including data at the event generator level (MC truth) and data mimicking the raw data format (digits), anchored to real data reproducing the running conditions;
- Event Summary Data (ESD) produced by the reconstruction algorithms, for both Monte Carlo and raw data. The ESD events provide calibrated tracks in a generic format, but also additional detector specific information allowing a full physics analysis;
- General purpose Analysis Object Data (AOD), derived from ESD data. The AOD data format contains a simplified event model with a few additional high-level detector specific parameters;
- Custom analysis object data, used standalone or together with the general purpose AOD for specific analysis;

- Published physics results and highly abstracted data resulting from the analysis; [2]

All these different kind of data deserve a PID. In case the logical datasets are split into multiple subsets for dissemination and management reasons, also these subsets may be identified by a PID.

3.1 Dynamic datasets

Identifying and providing persistent access to dynamic data can be challenging. Issues may arise where a researcher wishes to identify and reference: a) the subset of dynamic data used in research, b) the data used existing at a certain point in time [1]. There are some practices to address this issue:

- assigning a PID to each micro-subset of the data as it was retrieved at a given point in time;
- assigning a PID to snapshots and versions of the data at specified times or trigger events.
- assigning a PID to the query performed with the time-stamp

It is widely agreed that a one-size-fits-all approach does not currently exists to support all dynamic data citation use cases. However, a PIDs should be agnostic from the method used and should be flexible to support different cases.

3.2 Data versioning

There is currently no agreed standard or recommendation among data communities as to why, how and when data should be versioned. Some data providers may not retain a history of changes to a dataset, opting to make only the most recent version available. Other data providers have documented data versioning policies or guidelines based on their own discipline's practice, which may not be applicable to other disciplines. According to [1], there are two approaches to versioning: record-level and release-level that is more frequent in the life sciences domain. In fact, Some user communities need to resolve individual archived entities via a deterministically-versioned URI pattern, while others use the Release-level versioning for defined data releases.

There is currently a discussion in the global community as to the need for, or indeed the possibility of, an agreed best practice for data versioning across data communities. The Digital Curation Centre (DCC) recommends that data repositories should ensure that different versions are independently citable (with their own identifiers). However, this recommendation may not be applicable across data types and domains. For example, the Federation of Earth Science Information Partners (ESIP) specifies that a new DOI should be minted for each "major" but not "minor" version. However, a PID service should be flexible and accommodate every policy defined and recognized by the community of reference.

3.3 Complex data

Another challenge in identifying Big Data is related to their volume. The Open Data service provided by CERN is a valuable environment where the current PIDs system meets a challenge. Let us consider the dataset ATLAS ZPath 2015 Masterclass dataset 7 (http://opendata.cern.ch/record/359) composed of 20 files of ca. 30Mb each for 647.6 MB in total. Even if the record 359 is not assigned to a DOI yet, as soon as it will happen, the DOI will reference the landing page in which the 20 files, aggregated together by Content Management System on the base of the metadata, are displayed. How is it possible to verify if one of them changed or not What's happen if some of them, for any reasons, are moved to another location? And what about if the number of files composing a dataset is 4335 (e.g. dataset: MuOniaParked primary dataset in AOD format from Run of 2012, identified by the DOI: 10.7483/OPENDATA.CMS.ZCFQ.Q557 - 15.9 TB in total), but the relevant information has been discovered only in a subset of them?

3.4 Data location

The locations of the dataset on the Internet typically changing over time because of a number of reasons such as maintenance, business name of the organization changed, etc.. Thus, a reliable PID service needs continuous maintenance - an effort that is increasing with the advancement of e-Science and the advent of the Internet-of-Things (IoT). For instance, today billions of sensors and data sets are subject of PID assignment[50], but there is not any possibility to assume their URLs as a stable. In this respect, it is necessary that each time such kind of changes occur, every modification applied on the datasets is immediately reflected on the PID system. This implies the existence of a strong synchronization system between the PID system and the data centers, that intercept the relevant changes on the data side, and invoke the related update function (transactions) on PID system for alignment. At the moment, the current PID systems do not offer such kind of sync at all or in a way that can be easily embedded in the data center workflow.

3.5 Data delivery

It is well know that there is no consensus on what exactly a PID should point to.In [22], arguments to sustain the separation between the PID service and the data delivery are provided. In particular, the authors believe that making recommendations regarding the delivery of the information objects is outside the scope of a PID system. This is due to data services needed to provide their own interfaces and protocols for data delivery. Thus a PID should point to a landing page or an intermediate mechanism provided by the data sources. On the contrary, in [20], the access to the digital object is part of the PID service. In our perspective, a PID service should support both situations. Thus, a PID should manage the pointer to the dataset or to a content

negotiation layer by presenting a meta-resolution page in which are indicated (if any) multiple data access possibilities.

3.6 Data curation

The curation and preservation of scientific data are the main missions of the data archive such as DANS [17]. In principle a PID system and the data curation are decoupled services. Usually the PID systems do not ask for the identified data to be kept permanently. It is clear that such a condition affects the effectiveness of the PID in supporting reproducible research. In our perspective a trustworthy PID should identify only datasets that are curated according to standards such as ISO 14721:2012[18]. The results of experiments on Big Data curation presented in [36] and [46] are encouraging. In fact, the use of HDFS or IPFS to store the OAIS Archival Information Package (AIP) seems to be technologically feasible. In this respect, a T-PID should consider this aspect in the requirements.

4. Trusted Persistent Identifier system for Big Data

The idea to overcome the limitation of centralized or semi-distributed management adopted by the existing PID systems, has been explored in [12] [23] and then further elaborated in [30], [17]. In order to address the challenges introduced in section 3, in the present work we identify 5 pillars on which to build a trusted PID system (T-PID), as inspired by [9]:

- P1-Persistence of the community served. Despite the presence of several technological solutions, considering persistence as a purely technical problem is an understatement. In fact, persistence is deeply related with the interest and commitment that a particular user community has in preserving and making these resources accessible for future generations. In our vision, the concept of persistence in PID, should move from the persistence secured by the commitment of single institution/registration authority towards a shared commitment among the members of the user community served.
- P2-Persistence of the PID existence even if the resource is no longer available online(e.g., as proof that at some point the resource identified has existed).
- P3-Persistence of the PID resolution service. It is a commitment that ensures that the PID is resolvable to the dataset identified in the long term.
- P4-Persistence of the resource referenced. Ensuring long-term existence, accessibility and authenticity of the resource referenced by a PID needs to be considered an integral part of the T-PID.

[17] https://dans.knaw.nl/en
[18] Open Archival Information System https://www.iso.org/standard/57284.html

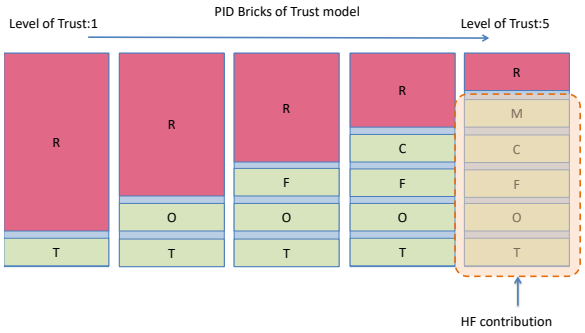

Figure 2: Bricks of Trust model

- P5-Persistence of the tracks of the changes occurred. In fact according to PREMIS recommendation [19], each update occurred on digital object at both metadata and data format, needs to be tracked. This should be true also for a PID, where a number of events may occur along its life-cycle. We reported here an example of events and the related process triggered in PID system that can benefit from the blockchain adoption.

All these requirements can be translated into a blockchain-based information system, following the guidance of the theoretical model named Mirror Model (MM). The Mirror Model [8] is a simple formal model that provides the general and formal conditions to yield trusted data and trusted procedures (that in the MM are mathematical functions). In MM, are identified three domains: Reality (where things objectively exist), Representation (a model conceived to depict the aspects of interest or concern in the Reality) and Trust (a structure of validation of the Representation of Reality whose aim is to make the Representation and its data dependable)[8]. In the T-PID context and thanks to the possibility offered by permissioned blockchain technology, Representation and Trust are merged together in a unique category, so that it is not necessary to mirror data, process and events into blockchain to obtain the desired level of Trust. In fact, the business logic of the T-PID can be entirely modeled within the blockchain environment (usually belonging to Mirror category) as demonstrated afterwards. Permissioned blockchain allows to provide benefits to all the Brick of Trust stack (BoT) through an all-in-one solution, as depicted in Figure 2.

In fact, at the technological level (T), the blockchain includes in a single and coherent infrastructure, a series of features that can be obtained only through various costly (F) ad-hoc implementations and the integration of different components, such as: permissioned membership (P1), immutability (P5), Performance, scalability, resilience (P2,P3), levels of trust (P4-partial), rich queries over an immutable distributed ledger, and so forth. Differently from public blockchain, permissioned blockchains maintain an access control layer to allow certain actions to be performed

[19] PREservation Metadata: Implementation Strategies - PREMIS https://www.loc.gov/standards/premis/

only by certain identifiable participants and allows role-limited implementations (e.g. allows peers to interact after a X.509 certificate-based authentication). This requires a organizational (O) model to define the underling business network properly. Moreover, the direct engagement of the members of the community in managing T-PID system (e.g. running a peer node) mitigates the risk of failure related to the commitment of a single organization, and set up the condition for the existence of a mandate that community gives to itself (M) to keep the service up and running in the long run.

4.1 Hyperledger Fabric

In order to implement the T-PIDs system we considered the use of Hyperledger Fabric (HF) [4] since allows a consortium-based blockchain implementation. HF introduces an execute-order-validate architecture[4] which is a fundamental shift from the traditional order-execute design followed by other blockchain platforms [5]. The goal is to separate transaction execution (via Smart Contract (SC)) from transaction ordering. All transactions submitted through a business network are stored on the blockchain ledger, and the current state of assets and participants are stored in the blockchain state database. The blockchain distributes the ledger and the state database across a set of peers and ensures that updates to the ledger and state database are consistent across all peers using a consensus algorithm. The consensus is achieved by having peer nodes in multiple organisations endorse transactions. Transactions are endorsed by executing chaincode, and signing the results of that execution. In order for the transaction to be committed by the blockchain network, all peer nodes endorsing the transaction must produce the same results from executing chaincode [20].

Compared to the traditional state machine replication approach [47], this architecture provides better scalability, new trust assumptions for transaction validation, support for non-deterministic SC, and modular consensus implementations. Thanks to these features, HF is being experimented in a very diverse contexts such as e-voting [10], IoT [19] or supply chain [43].

Chaincode (or Smart Contact) represent the business logic of the PID Business network and run concurrently in the network and can be deployed dynamically. It is a software in which are defined assets, participants and transaction instructions for manipulating the asset(s). Chaincode execution results in a set of key-value that are submitted to the network and applied to the ledger on all peers. Moreover, a smart contract enforces the rules for reading or altering key-value pairs or other state database information. Fabric allows parallel execution increasing overall performance and scale of the system. This first phase also eliminates any non-determinism, as inconsistent results can be filtered out before ordering. According to such features, HF has been considered suitable for T-PID implementation at functional level.

[20] https://hyperledger.github.io/composer/v0.19/integrating/call-out.html

4.2 T-PID Architecture

According to HF development practices, the fist step is the definition of a Business Network (BN) for T-PID management reflecting the actors involved, their roles and the relationships among them. In T-PID we can identify 2 roles: Consortium members running a peer node (C) and Users (U), so that the cardinality of the Actors (A) participating in the BN is given by $|A| = |C| + |U|$. There are no restrictions on the number of members composing P, so that it is possible to have $|A| = |C|$. However, there are technical considerations that needs to be done. In fact, HF is not based on Proof of Work (PoW) consensus mechanism, thus there is no need for all network participants to run a peer node to a act in the network. This condition allows infrastructure optimization and a better performance compared with permissionless solutions and it is reflected in the architecture here proposed for T-PID (see Figure 3).

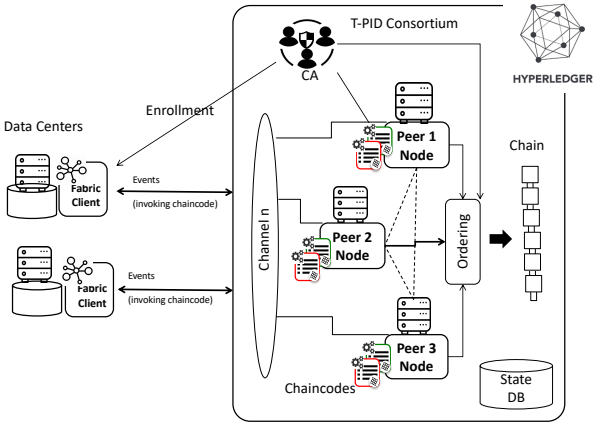

Figure 3: T-PID Architecture with IPFS node

The client node can be implemented as a plugin for the existing data management systems that interact through REST API of the chaincode in the blockchain or as a client tool able to detect new AIP published on the file system (e.g. HDFS) and interact with JSON RPC. In any case, this represents the gateway for the Users (U) to read/write on the blockchain.

Once the high level architecture is defined, the key elements to be identified and modeled in a HF based system are: Participants, Assets, Transaction and Event. In order to keep the example as simple as possible, we modeled the assets, transactions, and participants using the Hyperledger Composer notation language.

4.3 Assets

In HF, Assets can be tangible as well as intangible entities (e.g. goods, services, or property) and are stored in the ledger. Assets can represent almost anything in a business network and can be modelled through JSON-LD [21]. In this case, it is possible to use CouchDB as a state database, that allows complex rich queries against the chaincode data values, using the CouchDB JSON query language within chaincode. In our context the Assets identified are: Identifier (PID), PIDstate, and URIState and can be defined as follow:

```
namespace org.identifier
enum PIDState {
    o Active
    o Suspended
}
enum URLState {
    o Active
    o TmpUnavailable
    o Broken
}
asset Identifier identified by assetId {
  o String assetId
  --> Participant owner
  o String PID
  o PIDState state
  o String URLdataset
  o URLstate stateURI
  o String URLmetadata
  o String Hash
  o String ParentPID optional
}
```

In the T-PID, it is important to notices that the descriptive metadata are not included as a part of the PID asset definition. In this respect we manage two locators: "URLdataset" where it is possible to retrieve the data and the "URLmetadata" that points to a service in which it is possible to retrieve further information (e.g. landing page). Moreover the field "state" is used to manage the status of the PID (Active, Suspended), that can be set according to specific organizational rules; the filed "owner" represents the owner of the datasets and it is modelled as a Participant; the field "ParentPID" has been included to manage existing PID assigned to the entire set of datasets (e.g. DOI), while the field "Hash" is going to be used for the dataset validation check.

[21] https://json-ld.org/

4.4 Participants

The Participants set is mainly composed by Data centers curating scientific datasets and can be identified with simple information.

```
participant Actor identified by actorId {
  o String actorId
  o String Name
  o String VAT
  o String ....
}
```

4.5 Events

Events are defined in the business network definition in the same way as assets or participants. The event can be emitted by transaction processor functions to advice external systems that something happened to the ledger. Clients can subscribe to emitted events in order to react accordingly. The events matched with the transactions identified:

```
event PIDGeneratedEvent {
  --> Identifier asset
  o String PID
}
event URLUpdateEvent {
  --> Identifier asset
  o String oldURL
  o String newURL
  o String newHash
}
event NewURLAddedEvent {
  --> Identifier asset
  o String URL
}
event UpdatePIDMetadata {
  --> Identifier asset
  o String URL
  o URLState State
}
event CheckResultEvent {
  --> Identifier asset
  o String URL
  o URLState State
}
```

4.6 Transactions

Transactions are the mechanism by which participants interact with assets and a transaction may or may not lave a trace on the register. The transactions in a T-PID are several. In Figure 4, 5 relevant transactions have been identified.

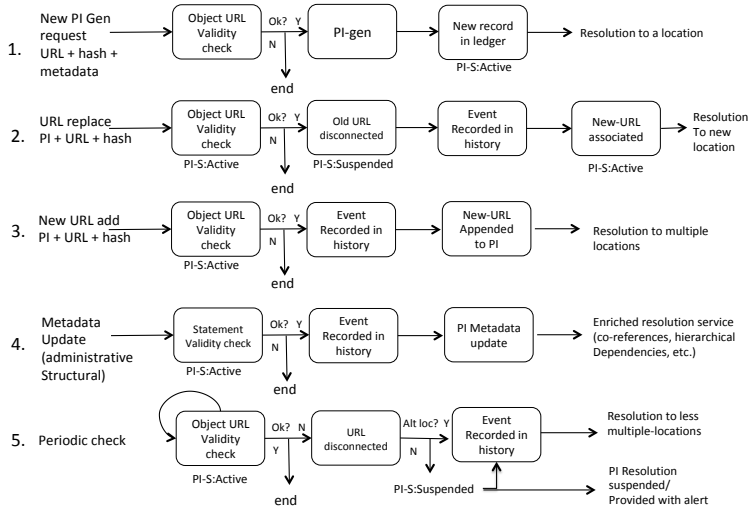

Figure 4: transactions

In order to demonstrate the functional feasibility of HF to support a T-PID implementation, we focus three of the most important transactions in a PID system: PID generation, PID resolution and PID Update.

4.6.1 PID Generation transaction

The transaction will be invoked by a specific HF client running on the data center side and the new PID is generated, recorded in the ledger and the related event generated.

```
/** transaction definition
transaction PIDAssignmentTransaction {
  o Identifier asset
}

/**
 * @param {org.identifier.PIDAssignmentTransaction} tx
 * @transaction
 */
async function PIDAssignmentTransaction(tx) {
```

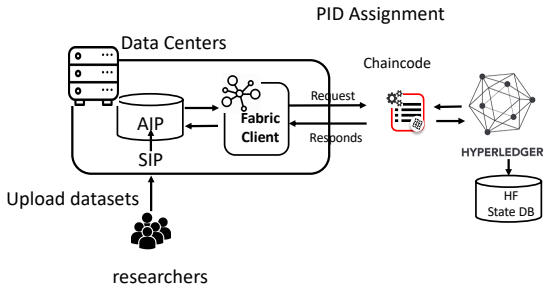

Figure 5: T-PID Assignment

```
return getAssetRegistry('org.identifier.Identifier')
.then(function(result) {
    var factory = getFactory();
    var newAsset = factory.newResource('org.identifier',
    'SampleAsset', tx.asset.assetId);
    newAsset.owner  =tx.asset.owner;
    newAsset.PID = //PID Generator
    newAsset.Hash=tx.asset.Hash
    // this can be imported from IPFS HASH
    //continue with property assignments
    let event = getFactory().newEvent('org.example.basic',
    'PIDGeneratedEvent');
 event.asset = tx.asset;
 event.PID=tx.asset.PID;
 emit(event);
    return result.add(newAsset);
 })};
```

4.6.2 PID Resolution

The Resolution service basically consists of extracting the URL associated to the given PID and sent it back to the users. To do so, a read-only transaction should be performed based on queries. Queries are an optional component of a business network definition, written in a single query file (queries.qry). In order to use the SQL-like notation of the query, it is necessary to implement the CouchDB in Fabric. The queries, even if are executed within a chaincode does not alter the ledger.

```
/** Sample queries to extract Identifier using PID
*/
query PIDResolution {
  description: "PID Resolution "
  statement:
```

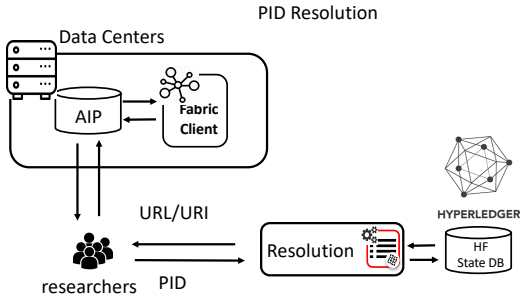

Figure 6: T-PID Resolution

```
      SELECT org.identifier.Identifier
          WHERE (PID==’_$PID’)
}

/** transaction definition
@commit(false)
@returns(String)
transaction ResolutionTransaction {

}

/**
 * @param {org.identifier.ResolutionTransaction} tx
 The sample resolution transaction instance.
 * @transaction
 */
async function ResolutionTransaction(tx) {

    // Get the asset registry for the asset.
    const assetRegistry = await
    getAssetRegistry(’org.identifier.Identifier’);
    // Search the asset in the asset registry.
    let result = await query (’PIDResolution’);
    if (result.lenght>0) {
    return result.URLdataset;
    }
}
```

4.6.3 PID Updates

Another example of implementation of a transaction is the management of dataset updates (e.g. relocation, versioning). The transaction receives as input the PID to be updated and the new URL and hash of the dataset. All these changes are recorded in the ledger, and permanently retained. In particular, in order to provide a dataset versioning service, HF enables information retrieval of the history of the updates occurred to a PID. In this way, the users can decide to use the value of the URL at a certain time in the past to access to a older version of a dataset. In this case both the value of the URL as well as the hash will be different. It is worth to remark, however, that the implementation of features like versioning, even if technical viable, are matter of the PID level of service defined.

```
transaction UpdateURLTransaction {
  --> Identifier asset
  o String newURL
  o String newHash
}

/**
 * @param {org.identifier.UpdateURLTransaction} tx
 The sample transaction instance.
 * @transaction
 */
async function UpdateURLTransaction(tx) {
    const oldURL = tx.asset.URL;
    const oldHash =tx.asset.Hash;
    // Update the asset with the new value.
    tx.asset.URL=tx.newURL;
    tx.asset.Hash=tx.newHash;
    tx.asset.stateURL='ACTIVE';
    // Get the asset registry for the asset.
    const assetRegistry = await
    getAssetRegistry('org.identifier.Identifier');
    // Update the asset in the asset registry.
    await assetRegistry.update(tx.asset);
    // Emit an event for the modified asset.
    let event = getFactory().newEvent('org.identifier',
    'UpdateEvent');
    event.asset = tx.asset;
    event.oldURL = oldURL;
    event.newURL = tx.newURL;
    event.newHash = tx.newHash;
    emit(event);
}
```

4.7 Ledger Storage management

In HF peers perform a versioning check against the transaction read set, to ensure data integrity and protect against threats. HF has concurrency control whereby transactions execute in parallel (by endorsers) to increase throughput, and upon commit (by all peers) each transaction is verified to ensure that no other transaction has modified data it has read. In other words, it ensures that the data that was read during chaincode execution has not changed since execution (endorsement) time, and therefore the execution results are still valid and can be committed to the ledger state database. If the data that was read has been changed by another transaction, then the transaction in the block is marked as invalid and is not applied to the ledger state database. However the invalid transaction are stored in the ledger. Peers cannot simply discard blocks and thereby prune Peer Ledger once they establish the corresponding vBlocks. To facilitate pruning of the Peer Ledger, a checkpointing mechanism has been introduced in HF 1.4. This mechanism establishes the validity of the vBlocks across the peer network and allows checkpointed vBlocks to replace the discarded Peer Ledger blocks. This, in turn, reduces storage space as well as the work to reconstruct the state for new peers that join the network.

5. Conclusion

This prototype, design an architecture and testing a subset of the core functionalities required for a PID (e.g. PID generation and resolution), demonstrate the feasibility of HF to be used to implement a T-PID systems. The proposed system can be evaluated according to 3 main criteria: Non-regression property, Rebound effect[11], Risk mitigation.

- a) Non-regression property: the validity of the development of the T-PID system is not affected by the addition of new elements. In fact the modification in business network or in business processes or in the data model does not invalidate the rest of the system requiring a total or partial rebuilding.
- b) Rebound effect [33]: the increased efficiency in managing technology complexity, business processes and trust creates an additional demand of new features and processes that can be managed at business logic (chaincode). It can be considered a proxy indicator of the achieved efficiency.
- c) The use of HF mitigates blockchain-based implementation risks (see Table: 1)

The use of HF to manage Big Data citation represents a promising opportunity to be further explored. In fact, the functioning logic, as well as the immutability and security of the technology, seems to fairly address all the PID system requirements. The possibility to manage the identification of different kind of Big Data with relative simplicity represents and added value for all the actors involved.

Risks of blockchain based project	HF based risk mitigation
failing in defining the technological requirements (risk associated with technology)	HF provides an all-in-one, modular-based, flexible, scalable, trustworthy and resilient infrastructure.
failing in predicting and adapting to the growth/evolution of the service demand with the result of having an over/under sized infrastructure (technology and financial risks)	HF infrastructure (e.g. number of peer nodes), can be tailored, within certain limits, without affecting its security and performance
adopting technologies that become obsolete in short time (risk associated with the adoption of standards)	the technology obsolescence is mitigated by the fact that HF is open source and supported by worldwide community leaded by top level industries such as IBM, Oracle, Microsoft and Accenture.
failing in defining the appropriate organizational model	HF allows a flexible and multi-role business network modelling. Changes in the business network can be easily deployed without affecting the validity of the transaction performed.
failing to determine the cost and sustainability model (technology and financial risks)	The shared approach allows a better sustainability for the T-PID system in the long run
failing in understanding and addressing community service requirements (risk associated with technology as well as community mandate)	The use of peer nodes as well as a closed memberships mechanism implies the existence of a strong mandate of the community to the peers that maintain up and running the infrastructure. The possibility to add new peers is always allowed.

Table 1: Blockchain Risks - HF-based Mitigation table

The possibility to design a distributed organization (Business Network) in which the actors can actually contribute to the infrastructure allows a better cost sharing and reinforce the commitment and mandate aspects for the reliability of the PID service. In turn, the impact for a decommissioning of an organization in the consortium is absorbed by the peers that can continue to maintain the PID system up and running. The possibility to manage multiple communities with different policies within the same consortium and blockchain implementation is another remarkable aspect to be considered. The use of a specific chaincode for each community as well as the possibility to exploit the channel feature of HF to implement a certain level of data privacy where necessary is definitively an element of flexibility that is currently missing in the current systems.

To conclude, even if blockchain technology seems to be structured to replace the

existing PID system infrastructures on which the current centralized (URN based), as well as the distributed (DOI) PIDs system, are built [41], the proposed approach is seen as an general improvement of the existing services because it addresses a not well-covered part of the pyramid presented in section 2. Thus, the mutual collaboration and possible future integration with the existing solutions is considered the best option to be considered.

References

[1] https://www.ands.org.au/working-with-data/citation-and-identifiers/data-citation/citing-dynamic-data

[2] ALICE Collaboration. ALICE data preservation strategy. CERN Open Data Portal, 2013 https://doi.org/10.7483/opendata.alice.54ne.x2ea

[3] Altman, M., Borgman, C., Crosas, M., Martone, M. An introduction to the joint principles for data citation. Bulletin of the Association for Information Science and Technology 41, 43-45 (2015).

[4] Androulaki E., Barger A., Bortnikov V., Cachin C., Christidis K., De Caro AS., Enyeart D., Ferris C., Laventman G., Manevich Y., Muralidharan S., Murthy C., Nguyen B., Sethi M., Singh G., Smith K., Sorniotti A., Stathakopoulou C., Vukolic M., Weed Cocco S., Yellick J., Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains, in EuroSys, 2018, pp. 30:1-30:15.

[5] Androulaki E., Cachin C., De Caro A., Sorniotti A., Marko Vukolic M., Permissioned Blockchains and Hyperledger Fabric ERCIM, 2017

[6] Baker, M., 1.500 scientists lift the lid on reproducibility. Nature News 533, 452-454 (2016).

[7] Beck K., Ritz R., Wittenburg P., Towards a Global Digital Object Cloud, Report from the Views on PID Systems training course and workshop In: RDA Europe Workshop August-September 2016, Max Planck Compute and Data Facility (MPCDF), Garching-Munich, German

[8] Bellini, A., Bellini, E., Gherardelli, M., Pirri, F., Enhancing IoT Data Dependability through a Blockchain Mirror Model, Future Internet, 2019, 11, 117.

[9] Bellini E., Bergamin G., Messina M., Cirinna' C., NBN:IT The Italian trusted persistent identifier infrastructure - Int.J. Knowledge and Learning, 2014, Vol 9, Issue 4.

[10] Bellini, E., Ceravolo, P., Damiani, E., Blockchain-based e-Vote-as-a-Service", IEEE 12th International Conference on Cloud Computing (CLOUD), 2019, 484-486

[11] Bellini, E., Coconea, L., Nesi, P., A Functional Resonance Analysis Method Driven Resilience Quantification for Socio-Technical Systems, IEEE Systems Journal, 2019, DOI:10.1109/JSYST.2019.2905713

[12] Bellini E., Damiani E., Fugazza C., Lunghi M., Semantics-Aware Resolution of Multi-part Persistent Indentifiers. WSKS (1), 2008: 413-422

[13] Bellini, E., Deussom, M. A., Nesi, P., Assessing Open Archive OAI-PMH implementations, Proceedings of the 16th International Conference on Distributed Multimedia Systems (DMS), 2010 pp. 153-158

[14] Bellini, E., Luddi, C., Cirinnà, C., Lunghi, M., Felicetti, A., Bazzanella, B., Bouquet, P., Interoperability knowledge base for persistent identifiers interoperability framework IEEE 8th International Conference on Signal Image Technology and Internet Based Systems (SITIS) 2012, pp. 868-875

[15] Bellini E., Nesi P., Metadata Quality Assessment Tool for Open Access Cultural Heritage Institutional Repositories. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7990 LNCS, 2013, pp. 90-103

[16] Bellini E., Nesi P., Coconea L., Gaitanidou E., Ferreira P., Simoes A., Candelieri A., Towards resilience operationalization in urban transport system, Risk Relibility and Safety: Innovating Theory and Practice (ESREL), 2016

[17] Bolikowski Ł., Nowiński A., Sylwestrzak W., A System for Distributed Minting and Management of Persistent Identifiers. International Journal of Digital Curation 10(1): 280-86, 2015

[18] Boulton, G., Reproducibility: International accord on open data. Nature 530, 281, 2016

[19] Brotsis, S. , Kolokotronis, N., Limniotis, K., Shiaeles, S., Kavallieros, D., Bellini, E., Pavué, C., Blockchain solutions for forensic evidence preservation in IoT environments", IEEE Conference on Network Softwarization (NetSoft),2019

[20] Bütikofer, N (2009). Catalogue of criteria for assessing the trustworthiness of PI systems, nestor-Materialien, Niedersächsische Staats und Universitätsbibliothek Göttingen In: Göttingen, German

[21] Cachin C., Architecture of the Hyperledger Blockchain Fabric, 2006

[22] Car N J.m Golodoniuc P., Klump J., The challenge of ensuring persistency of identifier systems in the world of ever-changing technology. Data Science Journal 16(13): 1-18,2017

[23] Ceravolo P., Bellini E., EPICA: Easy Persistent Identifier Common Architecture, In: Meersman R., Dillon T., Herrero P. (eds) On the Move to Meaningful Internet Systems: OTM 2010 Workshops. OTM 2010. Lecture Notes in Computer Science, vol 6428. Springer, Berlin, Heidelberg

[24] Ceravolo, P., Bellini, E., Towards Configurable Composite Data Quality Assessment, IEEE 21st Conference on Business Informatics (CBI) 1, 249-257, 2019

[25] Chard K., D'Arcy M., Heavner, B., Foster I.,Kesselman C., Madduri R., Rodriguez A., Soiland-Reyes S., Goble C., Clark, K., Deutsch E. W., Dinov I., Price N., Toga A., I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets, 2016 IEEE International Conference on Big Data (Big Data), IEEE, pp. 319-328

[26] Chen X., Dallmeier-Tiessen S., Dasler R., Feger S., Fokianos P., Gonzalez J.B., Hirvonsalo H., Kousidis D., Lavasa A., Mele S., Rodriguez D.R., Šimko T., Smith T., Trisovic A., Trzcinska A., Tsanaktsidis I., Zimmermann M., Cranmer K., Heinrich L., Watts G., Hildreth M., Lloret Iglesias L., Lassila-Perini, K., Neubert S. , Open is not enough, Nature Physics,2018 https://doi.org/10.1038/s41567-018-0342-2

[27] Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. FORCE11 doi:10.25490/a97f-egyk (2014).

[28] Duerr R. E., Downs R. R., Tilmes C., Barkstrom, B., Lenhardt W. C., Glassy J., Bermudez L. E., Slaughter P., On the utility of identification schemes for digital earth science data: an assessment and recommendations. Earth Science Informatics, 2011, 4:139-160,

[29] Ferguson C., McEntrye J., Bunakov V,, Lambert S., Sandt S., Kotarski R., McCafferty S., D3.1 Survey of Current PID Services Landscape (Version 1), FREYA project, 2018

[30] Golodoniuc P., Car N. J., Klump J., Distributed Persistent Identifiers System Design, Data Science Journal, 2017

[31] Golodoniuc P., Car N. J., Cox S. J. D., Atkinson R. A., PID Service an advanced persistent identifier management service for the Semantic Web, 21st International Congress on Modelling and Simulation, 2015

[32] Goodman, S. N., Fanelli D., Ioannidis J. P. A., What does research reproducibility mean?, Sci. Transl. Med. 8, 341ps12 (2016).

[33] Gossart C.,Rebound effects and ICT: A review of the literature - in Ed.: Hilty, Aebischer - ICT Innovations for Sustainability, Springer International, 2014, DOI:10.13140/RG.2.1.3301.3926

[34] Huber R., Klump J., How dead is dead in the PID Zombie Zoo?, In: RDA Europe Workshop August-September 2016, Max Planck Compute and Data Facility (MPCDF), Garching-Munich, Germany

[35] Jacobson V., Smetters D. K., Thornton J. D., Plass M. F., Briggs N. H., Braynard R. L. , Networking named content, in Proceedings of the 5th international conference on Emerging networking experiments and technologies. Rome, Italy: ACM Press, Dec. 2009. doi:10.1145/1658939.1658941 p. 1.

[36] Krisostomus Nova, Rahmanto, Mardhani Riasetiawan, Data Preservation Process in Big Data Environment using Open Archival Information System, 4th International Conference on Science and Technology (ICST), 2018, Yogyakarta, Indonesia

[37] Kunze, J. The ARK Persistent Identifier Scheme. Internet Draft, 2007. http://tools.ietf.org/html/draft-kunze-ark-14.

[38] Loewenstern A., Norberg A., The BitTorrent Protocol Specification- BEP 5, Mar. 2013. [Online]. Available: http://www.bittorrent.org/beps/bep 0005.html

[39] Maymounkov P., Mazi'eres D., Kademlia: A Peer-to-Peer Information System Based on the XOR Metric,in Peer-to-Peer Systems. Berlin, Heidelberg: Springer, 2002, vol. 2429, pp. 53-65.

[40] McMurry JA, Juty N, Blomberg N, et al., Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol. 2017;15(6):e2001414. Published 2017 Jun 29. doi:10.1371/journal.pbio.2001414

[41] Mirek Sopek, Grzegorz Zycinsky, Using Blockchain for Digital Identifiers: Improving Data Security and Persistence for Digital Object Identifier (DOI) and Legal Entity Identifier (LEI), The E-Finance Lab and DZ BANK 2016 Fall Conference, Goethe University Frankfurt. September 1st, 2016.

[42] Paskin, N., Digital Object Identifiers. Inf. Serv. Use, 22(2-3):97-112, 2002

[43] Perboli, G., Musso, S., Rosano, M., Blockchain in logistics and Supply Chain: A Lean Approch for Designing real-World Use Cases", IEEE Access, 2018 DOI:10.1109/ACCESS.2018.2875782

[44] Sam X. Sun. Internationalization of the Handle System - A persistent Global Name Service. 1998

[45] Sarala M. Wimalaratne, Nick Juty, John Kunze, Greg Janée, Julie A. McMurry, Niall Beard, Rafael Jimenez, Jeffrey S. Grethe, Henning Hermjakob, Maryann E. Martone, Tim Clark- Uniform resolution of compact identifiers for biomedical data - Scientific Data volume 5, Article number: 180029, 2018

[46] Sawood Alam, Mat Kelly, Michael L. Nelson (2016) Interplanetary Wayback: The permanent web archive, IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2016, DOI:10.1145/2910896.2925467

[47] Schneider B., Implementing Fault-tolerant Services using the State Machine Approach: A Tutorial, ACM Comput. Surv., vol. 22, no. 4, pp. 299-319, Dec. 1990.

[48] Sicilia M.A., García-Barriocanal E., Sánchez-Alonso S., Cuadrado J.J., Decentralized Persistent Identifiers: a basic model for immutable handlers, Procedia Computer Science, 2019, DOI:10.1016/j.procs.2019.01.087

[49] Sukhwani H., Wang N., Trivedi K.S., Rindos A., Performance Modeling of Hyperledger Fabric (Permissioned Blockchain Network) IEEE 17th International Symposium on Network Computing and Applications, 2018

[50] Wannenwetsch O., Majchrzak T.A., On constructing persistent identifiers with persistent resolution targets - 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pages 1031-1040, 2016

[51] Wimalaratne S. M., Juty N., Kunze J., Janée G., McMurry J. A., Beard N., Jimenez R., Grethe J. S., Hermjakob H., Martone M. E., Clark, T., Uniform resolution of compact identifiers for biomedical data, Scientific Data 5, Article number: 180029, 2018

Received 6.02.2019, Accepted 17.09.2019