DisTRi: A Distributed Trusted Rights Framework for Digital Content

Authors: ***A. Galuten, P. Jessop, J. Lacy, D. Maher, T. Pan, P. Rixhon, M. Simpkins***

1 Abstract

The authors assemble new, breakthrough technologies that focus on the fundamentals of rights management, namely content identifiers, stakeholder identifiers, metadata associations, authoritative assertions, and the use of trusted, multi-party, distributed, dynamic data management systems to create and share Rights Management Information. By providing trustworthy Rights Information, the provenance, authenticity, and compliance with agreements enable the automated distribution of content and associated rights compensation. The approach is minimally prescriptive but maximally supportive and inclusive. It allows many solutions to be used while enabling numerous ways in which individuals and organisations can cooperate in originating, enriching, governing, and distributing trusted Information, helping streamline current processes and trigger innovative businesses.

Keywords: Creative Industries, Distributed Ledger Technologies, Content Identification, Rights Management Information, Copyright, Attribution, Data Cooperatives, Data Governance, Data Provenance

2 Introduction

Recent advances in various areas of computer science and systems allow us to credibly assemble a truly useful global Rights Framework for digital content. In this paper, we show that it is possible to provide an extensible set of software components and tools that allow individuals and other entities to make useful and authoritative assertions about content, and allow others to easily discover and rely on (i.e. trust) those assertions for any useful and ethical purpose, including especially the automation of attribution, distribution, remuneration, etc. We explain the goals of the Framework, its important properties, and how it is initially configured to immediately provide for a growing copyright infrastructure enabling a vibrant global market for any form of digital content. We then describe a Framework that is complete enough to enable the near-term needs of existing media markets and minimally restrictive and extensible enough to accommodate new markets and new automated approaches to content discovery and distribution that rely on provenance, authenticity, and other properties about content and associated metadata.

This Framework – the Distributed Trusted Rights Framework (DisTRi) – is a structure underlying the exchange of Rights Management Information that is built upon a set of foundational standards and technologies supporting a distributed network of copyrights declarations, attributions, attestations, and queries for the digital era. DisTRi maintains a set of functions and application programming interfaces (APIs) that enable creators and other stakeholders to register

works and related subject matter such that the works, associated metadata, and creators' identities are immutably bound, the binding time-stamped and registered within the Framework. DisTRi will be distributed among multiple independently operated services that can use capabilities and resources from one another to implement the DisTRi APIs and functions.

New technologies exploited in this Framework include:

- Open content identification – new methods and services that identify content at the time of creation and bind those identifiers to the content itself
- Means of immutably binding content metadata, that nonetheless allow for updates and revocation
- Means for search and discovery with authoritative, authenticated response
- Use of self-sovereign identity and identity management technologies that help ensure authority, confidentiality, and privacy of various aspects of content management and distribution.
- Human-centric approach – though we aim at enabling massive automation, people will be providing the information and relying on it. We emphasise tools that ensure that non-technical people can understand and use the essential instruments that can affect them
- Use of new distributed systems technology that ensure scalability and efficiency of services that can be built using the Framework
- Use of new trust management technologies and techniques that ensure the reliability of the data.

3 Objectives of the Rights Framework

DisTRi is the engine of the Internet of Value to power the media industry [1]. Copyright is a by-product of content creation that becomes the creator's currency. Hence, DisTRi will be the keystone of a clearing system handling hundreds of billions of Euros. The Framework will also be used to improve the certainty around the provenance of media and combat fake news [2].

The Rights Framework supports trusted attestations of attributions of rights to rightsholders. It answers the questions "who did what" and "who owns what". It helps protect authors' moral rights – allowing them to publicly declare the rules for the use of their work and their work to be correctly credited – as well as rightsholders' material rights – enabling the fair, appropriate, proportionate, and transparent remuneration of usage. The Framework copes with single or collaborative creations, and with static or dynamic productions.

DisTRi is an open Framework that creates a public and auditable register of Rights Management Information, hereby realising the potential of new policymaking initiatives such as the European directive on copyright [3], the Music Modernization Act [4] in the US, or the European strategy for data [5]. It fosters efficient B2B and B2C Electronic Markets for Media Assets and paves the way for new business models such as micro-licensing and secondary markets. Then, it protects

moral and material rights in User Generated Content – not only the rights of re-used original work but also the rights of newly generated creations. Finally, it helps prevent the misappropriation of creative content.

3.1 Qualities of enabled systems and methods

What are the qualities of the systems and methods enabled by DisTRi?

Open standards – DisTRi relies on open specifications, standardised Rights Management Information and data exchange standards [6]. They encourage stakeholders' collaboration and adoption, and help them adapt their business models, create, and offer new services. They also ensure transparency, interoperability, and data normalisation across the Framework. The user community, united around improving the infrastructure, co-specifies swiftly reliable, secure, and cost-effective concepts and capabilities. The users of the DisTRi Framework will not be locked in by one vendor or one technology.

Inclusive architecture – The DisTRi Framework adopts an inclusive architecture. It will be used by multiple stakeholders of the value network – authors, rightsholders, Collective Management Organisations, information society service providers and users. It will register single or collections of rights declarations. A multitude of content creators, users, prosumers, organisations, and marketplaces will be able to query the Framework and access Rights Management Information. They will also be able to build services and applications either to feed the Framework with declarations and metadata or to leverage the Rights Management Information for commercial or not-for-profit purposes.

Proven technologies – DisTRi depends on the distributed consolidation of concepts that are proven technically and legally. It anticipates future developments by monitoring technical advancements, changing circumstances of the creative industries, and progress of policymaking. Various methods and tools are combined to address specific issues, integrate the solution, and ensure backwards compatibility for fast mass adoption.

Scalability – The DisTRi Framework is a network of trusted ledgers explicitly designed to scale by virtue of its distribution across numerous dimensions: physical, sectorial, and jurisdictional. Rather than being one huge ledger with large numbers of redundant nodes, it is a network of many smaller ledgers, each of which specialises in specific rights domains with their particular policies complying with their local jurisdictions.

Sustainability – DisTRi uses a meshing strategy that takes advantage of the fact that over time thousands of Trusted Immutable Distributed Assertion Ledgers (TIDALs) will emerge. The meshing strategy will allow all ledgers to help check one another while conserving resources and obviating the need to use expensive proof-of-work and similar energy consuming protocols.

Portability – The DisTRi Framework is aligned with the GAIA-X project developing the foundations for a federated, open data infrastructure, connecting centralised and decentralised

infrastructures to turn them into a homogeneous, user-friendly system [7]. The resulting federated form of data infrastructure strengthens the ability to both access and share data securely and confidently – respecting both individual privacy and business confidentiality.

Decentralisation and other key principles for social good – DisTRi will not only transform technical structures, but it will also transform organisational and financial dimensions of the creative industries towards more equitable network governance or distribution of value and resources [8]. The DisTRi Framework will capitalise on and live up to the principles of Distributed Ledger Technologies: technological decentralisation, governance decentralisation, autonomy, openness, privacy, and a data commons approach emphasising the relational production, shared value, and collective governance of data. DisTRi supports multiple interoperable policies, be they defined by a specific creative sector or a particular jurisdiction – national, regional, or global.

3.2 Architectural overview

All the technologies we are proposing are designed to be open, available to all and designed to facilitate compatibility among multiple parties. The Framework is architected to support virtually all creative industries and their associated content formats. Thanks to the open, standards-based approach we are taking, DisTRi is easy to access, and will enable individuals and small companies to create viable and competitive businesses. It will also allow established stakeholders to streamline their processes.

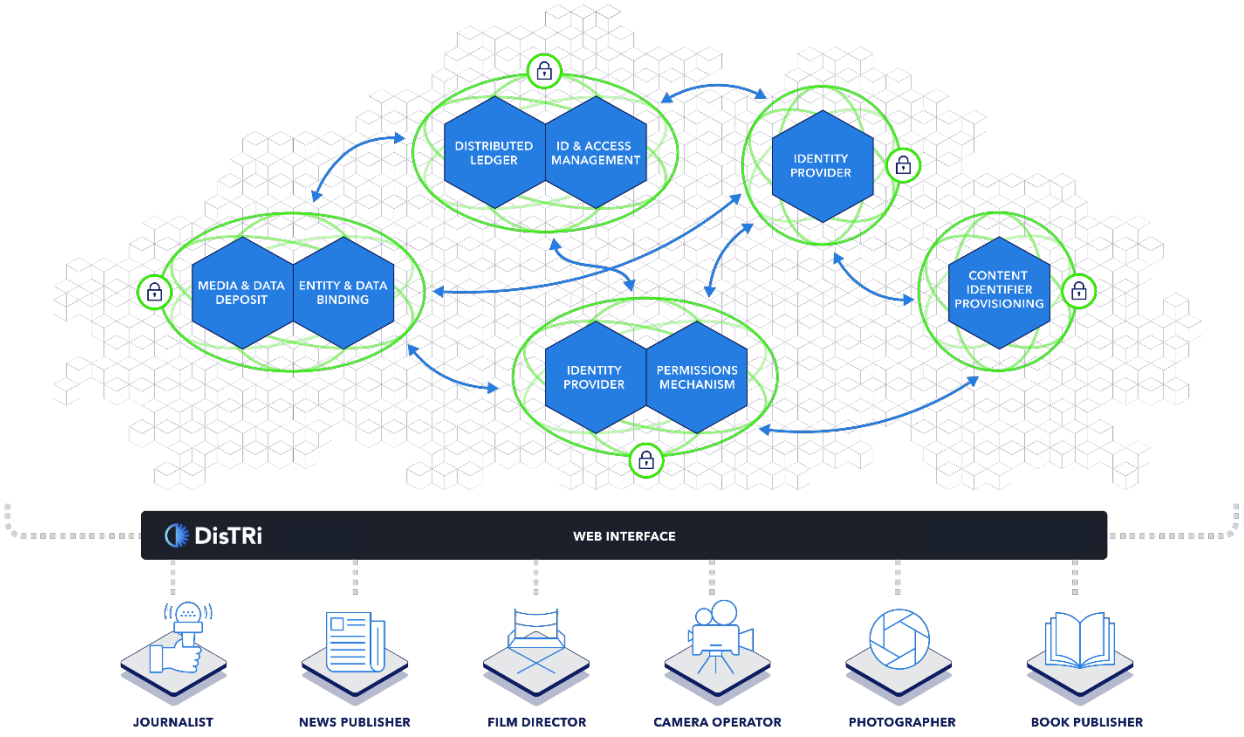

Figure 1: High-level architecture of the DisTRi Framework

Users access DisTRi through a Web interface, but they could also use applications that communicate with the Framework. In the same way that today's Web pages combine elements from different servers, our Web interface (or third-party applications) can access different elements from multiple parties that will all work together seamlessly.

Looking at the architecture depicted in Figure 1, there are common elements that apply to any creative ecosystem.

First there are the entities and individuals that contribute to the DisTRi Framework, shown here across the bottom of the drawing. These entities are a small representation of the numerous stakeholders we can support from music, film, journalism, book publishing or other creative sectors.

If authors and performers want to be credited and remunerated for their creative works, they must be identified. DisTRi will support use of any standards-based identification services, including Creative Passport, Self-Sovereign Identity, and others. We describe these services in more detail later in this paper.

In the DisTRi Framework, identity and access services allow users to manage their works and share them with others throughout the content value network. These services also tell DisTRi who should be credited and remunerated. If a journalist writes an article and wants to send it to one publisher or twenty, that is the author's prerogative and should be under the author's control. And, if the author and those publishers agree on terms of remuneration, any payment system will be able to query the DisTRi Framework to trace the author's contribution.

There are a few other elements needed to make the Rights Framework robust.

1. The data representing the creative elements themselves must be stored. In DisTRi, data can be stored in the cloud or locally.

2. A ledger must be used to maintain records of all transactions. These records must be immutable so that their veracity can be trusted. Distributed Ledger Technology is the natural choice. To mitigate issues with power consumption, latency and cost, the DisTRi Framework uses particularly efficient and scalable TIDALs to maintain these records.

3. One further requirement for a complete and robust Framework is a permissions mechanism. Our permissions mechanism defines the actions that each user can perform in the Framework, based on that user's identity and associated permissions. In our demonstration, we are using a platform that meets our security and scalability requirements. It has been proven effective in several existing markets including national power companies and large social networks. Ultimately, as we seek to standardise the solution, it is the requirements that will drive the standard, not the specific implementation.

4. One of the key elements of the DisTRi Framework is content identifiers. Content identifiers are unique ID codes for creations. In the same way that books rely on ISBN numbers, digital content (which we understand here to mean content in a digital encoding following known rules) uses universal identifiers for generic media-types such

as text, image, audio, and video. There are many content identifiers in use today, but they are not well suited for an open distributed system. One of the main problems is that they are not generated when the work is created. Instead, they are generated after the work has passed through multiple hands. Therefore, they fail to account for all the creative contributions to the work. In DisTRi, we use a number of content identifiers called Identity Bindings, or IDBs for short, including the new International Standard Content Code, referred to as ISCC, which proposes a universal and decentralised identifier for texts, images, audio, and videos.

When a new creation is first deposited into the Framework, content identifiers are generated. The creation is timestamped, signed, linked to associated works and hashed so that its provenance is clear. This content identifier is linked to it for its lifetime. For example, if a textual work is transformed into a movie or a composition is performed and recorded, new content identifiers are created which are bound along with the original identifiers so that the original creator can be credited and compensated. This will allow us to prevent attribution errors from occurring rather than building complex systems to correct them after the fact. This chain of handling and control also means that, in the event of disputes, all provenances can be verified.

All these components together comprise the DisTRi Framework: identity and access management, content identifier generation and management, data storage, distributed ledgers, permissions mechanism, and a secure execution environment.

After having described the components of the DisTRi Framework and their interrelationships in greater detail in the following paragraphs, we will discuss its extensibility and market sensitivity, touching upstream functionalities such as the ingestion of metadata (discussed below) pertaining to new creations, catalogues, cultural heritage, or orphan works, and downstream functionalities such as content licensing and distribution, usage monitoring, and royalty payments.

4 Works and Rights Management Information

Frameworks within which entities can be described are remarkably resilient to differences in the purpose for which the entity is created. The motivations and experiences of a scientist, a novelist, a composer, a pamphleteer, a financial analyst, and a coder are very different. But all start with an idea that exists only in their head, finds some expression, and is recorded in a format that can be shared with others.

While librarians developed a framework under the title of *Functional Requirements for Bibliographic Records* (FRBR) [9], another group centred around publishing developed the <indecs> framework [10] which is at first glance very similar but in fact takes a rather different, and arguably more robust, approach. Each recognises that the various forms in which creations are found (as a consequence of their stage of creation) are related one to another and require coherent approaches to their description. This can lead to them being useful to their audience by being findable and referenceable. The use of the term "metadata" is customary for the data effecting this description even though this use is sometimes inconsistent with the customary definition of that term: "data about data."

In practice, it is usually not sufficient to simply describe an entity, it is necessary to be able to use a unique, codified identifier to stand for the entity so that it can be cited without ambiguity. These identifiers are familiar in everyday life: car registration plates, passports and social security numbers, product barcodes, etc. Such identifiers lubricate systems but are in practice of little value unless there is trust in a mechanism to link the identified entity with the identifying code. Traditionally (though other approaches are described elsewhere in this paper) this has taken the form of a registry held by someone who is trusted by the users of the system. This contains the codes and information about the identified entity. For example, a government department holds a list of car registration plates together with information about the car: its model, chassis number, colour etc. Because people trust that the government will not capriciously alter this registry and there are criminal penalties for carrying a false plate, users are confident that a car bearing a particular plate has the registration number it carries because it can be checked in the registry.

This approach of linking or "binding" entities and their identifiers leads to a useful classification of metadata relating to an entity:

> *Reference metadata* (sometimes called minimum metadata or kernel metadata) is the smallest set of data which uniquely defines an identified entity. Although different such sets can be identified, there is usually community consensus on the appropriate set to use in particular circumstances.
>
> *Rights management metadata or attribution metadata* is the information relating to the ownership, use and reuse of a created entity, and
>
> *Indexing metadata, rich metadata or discovery metadata* is essentially everything else relating to the identified entity. This can include information helpful in searching for a particular entity (the genre of a novel or musical composition), for aiding its enjoyment or understanding (the influences on the creator or School of creators) or for general interest (the location of the first performance of a concerto).

All this data has provenance – who asserted the information and when. This is sometimes regarded as the fourth category: *administrative metadata.* It can be important in assessing the reliability, value, and trustworthiness of the other data.

As mentioned above, all these types of data have utility whether the identified entity was created for science, literature, art, politics, commerce or engineering. They apply to the scientific journal article, the novel, the concerto, the tract, the report, and the software module. However, where the creation is protected by copyright (which would usually include all of the above examples) the data may be considered special in the terms of the 1996 WIPO treaties [11] – the WIPO Copyright Treaty and the WIPO Performances and Phonograms Treaty. This pair of treaties defines and protects a class of metadata called Rights Management Information (RMI). Both treaties define RMI as

> *information which identifies the work, the author of the work, the owner of any right in the work, or information about the terms and conditions of use of the work, and any numbers or codes that represent such information, when any of these items of information*

is attached to a copy of a work or appears in connection with the communication of a work to the public.

The European directive 2001/29 [12] on the harmonisation of copyright in the information society expands the WIPO definition of RMI and stipulates that

> *"Rights Management Information means any information provided by rightsholders which identifies the work or other subject matter [ ], the author or any other rightsholder, or information about the terms and conditions of use of the work or other subject-matter, and any numbers or codes that represent such information. [This] shall apply when any of these items of information is associated with a copy of, or appears in connection with the communication to the public of, a work or other subject matter [ ]".*

There is a broad consistency between the definitions of RMI and the definition of rights management data above. Information about terms and conditions for the use of the work is seldom communicated technically beyond blunt statements of permission or prohibition ("all rights reserved") but where these expressions are attached to work or communicated in conjunction with it, they attract the protection specified in the treaties.

RMI are not only defined, but they also become more and more regulated. The European Union continued to adapt its copyright policies to the digital era and issued in 2014 a directive [13] which, with respect to musical work licensing, requires that

> *"[ ] a Collective Management Organisation which grants multi-territorial licences for online rights in musical works has sufficient capacity to process electronically, in an efficient and transparent manner, data needed for the administration of such licences, including for the purposes of identifying the repertoire and monitoring its use, invoicing users, collecting rights revenue and distributing amounts due to rightsholders.*
>
> *[ ] a Collective Management Organisation shall comply, at least, with the following conditions: (a) to have the ability to identify accurately the musical works, wholly or in part [ ]; (b) to have the ability to identify accurately, wholly or in part, with respect to each relevant territory, the rights and their corresponding rightsholders for each musical work or share therein [ ]; (c) to make use of unique identifiers in order to identify rightsholders and musical works, taking into account, [ ] industry standards and practices [ ]; (d) to make use of adequate means in order to identify and resolve in a timely and effective manner inconsistencies in data held by other Collective Management Organisations [ ]".*

Similarly, the US Music Modernization Act [14] foresees that

> *"The mechanical licensing collective is authorized to [ ] maintain a publicly accessible database of musical works (and shares of such works) and copyright owners, and other information relevant to the administration of licensing activities [ ]".*

Practically, the reliability of the binding between an identifier and the identified entity has traditionally depended on the governance of the identifier – the processes adopted to define the fine grain "rules of the road" within the broad confines of the specification and to operate any

required infrastructure. This has also determined certain parameters such as the cost of minting a new identifier and the costs of verifying it and looking up any metadata held within the system.

The binding was introduced above and will be developed in the next section. This binding has a number of dimensions which include cost (is it expensive to assert the binding or to verify it?), reliability (is it possible that the binding connects an identifying code to the "wrong" entity?) and persistence. Persistence itself has several aspects summarised [15] by Andrew Treloar with respect to "objects" (though abstract and personal entities are also relevant):

- Persistence of object (or a mechanism to handle its non-persistence)
- Persistence of identifier
- Persistence of binding between identifier and object
- Persistence of service to resolve from identifier to object
- Persistence of service to allow for updating of binding between identifier and object

For the purposes of rights management, it might be argued that persistence is not required beyond the longest surviving right (say 170 years – the life of an infant author plus 70 years) when the rights fall into the public domain but in practice these systems have other uses and longer persistence may be required.

And the persistence is not just a matter of retaining databases – there is also the matter of technology and business incentives. Even though simple technologies like cryptographic hashes (which change when a single bit is altered) will likely persist for a long time, the business reasons to maintain the data associated with those hashes may not. As long as the data structures are distributed and interoperable, they will be kept alive by the interested parties.

5 Content identifiers

As of July 2020, approximately 59% of the world population (i.e., 4.57 billion people) are actively using the Internet [16]. The sheer amount and transformative impact of digital content demands for a review of our current content management methods.

A crucial prerequisite to operating efficiently in this challenging environment is the ability to identify and reference digital content swiftly, reliably, and independently of its storage location. The structure and management of global content identifiers strongly correlate with the grade of achievable automation and the potential for innovation within and across numerous industries.

We will first list the content identification requirements for the DisTRi Framework and then follow with a discussion of some existing and future content identification technologies and how these technologies meet our requirements.

5.1 DisTRi content identification requirements

Based on the objectives of the DisTRi Framework, we define the following requirements for content identification:

Legacy Support - International standard identifiers like ISBN[1], ISWC[2], ISRC[3] and others identify entities: creations, works, recording events or products and play essential and specialised roles in several existing industries and must continue to be supported in order to accommodate content ecosystems that rely on them.

Decentralization - Content identification should have a low barrier to entry. By using a system where the creation and assignment of content identifiers to digital assets does not depend on a centralised service, these barriers are minimised.

Universality - The system should be generic and support all media types across all creative sectors.

Extensibility - The identifier system shall be extensible such that it can adapt to a broad range of use cases.

Automation - The identifier system should minimize the requirement for human-curated input.

Irremovable - Many platforms automatically remove embedded metadata from digital assets. Identification should not depend on having to carry along a content identifier together with a digital asset.

Verifiability - The binding between an identifier and the associated digital asset should be independently verifiable.

Low Overhead - The system should have a low administrative overhead and it should be inexpensive to acquire an identifier for digital content such that associated information can be managed from the very beginning.

Inclusiveness - The identifier system must be non-proprietary and openly documented such that ecosystem participants do not depend on a single implementation of the system.

Robustness - The identifier system should account for the dynamic nature of digital content and support automated detection and matching of related content.

5.2 Content identification technologies

To meet these and other requirements, the DisTRi Framework uses a combination of the following three different approaches to content identification, each of which serves a different set of purposes.

1. Existing legacy identifiers
2. Cryptographic hashes
3. ISCCs[4]

[1] ISO 2108 International Standard Book Number (see https://www.isbn-international.org/content/what-isbn)
[2] ISO 15707 International Standard Musical Work Code (see https://www.iswc.org/what-iswc)
[3] ISO 3901 International Standard Recording Code (see https://isrc.ifpi.org/en/)
[4] International Standard Content Code (described in detail in section 5.2.4)

To support a digital economy, we need to bind the identity of the author of the content with their creative work. We call this binding an Identity/Identifier Binding (IDB) where these elements and the content that they reference are bound cryptographically, and that binding is registered on a distributed ledger as discussed later.

A typical such binding for Content X with Metadata M on behalf of creator with identity represented by CID would look something like:

IDB = Hash (Hash (X), M, CID)

IDB would then be registered in a distributed ledger. If new metadata were to be added (e.g., a legacy identifier or an ISCC), a new ledger entry would be created as follows:

IDB-update = Hash (IDB, ISCC, Legacy ID, CID)

The IDB-update would also be registered with the distributed ledger. (See the discussion later in this paper on TIDALs and TIDAL derivatives for more on this process). At this point, IDB can be used to reference the original content and all metadata or content related to it.

More detail about legacy identifiers, hashes, and ISCC is provided in the following sections.

5.2.1 Legacy identifiers

International standard identifiers like ISBN, ISWC, ISRC and others approach the content identification problem by using human-curated descriptive metadata as a proxy for the referenced abstract content, together with access-controlled and registry-based systems.

These systems usually identify entities like creations, works, recording events or products and play essential and specialised roles in several industries. Because these industries are existing businesses with multiple value chain participants, we must continue to support such legacy identifiers to accommodate content ecosystems that rely on them. As stated above, one way that the DisTRi IDB approach handles these types of identifiers is to treat them as additional metadata to be bound to a manifestation (e.g., a hash) of the creative work via cryptographic hashes.

It should be noted however, that these legacy approaches do not provide universal data-driven identifiers for digital manifestations of content (data) and communities with large amounts of dynamic, granular and short-lived digital content (e.g., journalism, photography, user-generated content) have not yet adopted interoperable standard identifiers.

The existing identifiers also do not directly target the generic problem of content-based identification (nor were they designed for this). Consequently, they do not answer questions like: "Given some digital manifestation of content (data), what are the identifiers that I can use to find metadata or reference the content?"

5.2.2 Cryptographic hashes

Another approach used in technology-driven contexts is the use of cryptographic hashes. With cryptographic hashes, precise identification of data is a simple, automated, and reliable task.[5]

Association of time, dates, digital signatures, and cryptographic hashes with creative works must be supported to provide an absolute (non-fuzzy) attestation of a specific creation at the time of creation. To resolve a copyright dispute, there must be a proof of a specific deposit at a specific time and date by a specific person.

DisTRi accomplishes this by cryptographically binding a manifestation of the creative work (e.g., a cryptographic hash of the work) itself to the identity of the creator using standard cryptographic hash functions and protocols (e.g., SHA-256[6]) and standard digital signatures (e.g., RSA PKCS#1_v1_5[7]). These hashes and signatures are created at the time of deposit of the work and are time stamped to avoid any ambiguity about provenance. These IDBs can be updated as additional metadata is collected. For example, when a legacy identifier (like ISRC, ISWC, etc.) is associated with the creative work, it can be added, and the new collection of data can be signed and hashed attesting to the binding. In this way, any of the identifiers can be used to find any of the other identifiers for resolution, attribution, or payment.

It is important to bear in mind that cryptographic hashes are good at identifying static data, but they are not good at identifying re-encoded, resized, or re-compressed content. Every little change in the data that encodes the content results in a completely uncorrelated cryptographic hash, even if the perceptual content remains unaltered. In this case it is then very difficult to resolve the metadata unless our identifiers are still embedded in the digital asset. In cases where the media has been changed and the underlying identifiers are no longer attached, we must rely on content recognition systems.

5.2.3 Content recognition

Content recognition is a notoriously hard problem. In Jinda-Apiraksa et al., 2013 [17], the authors tasked ten human subjects to review 701 photos and identify near-duplicates. The results illustrate the nature of the problem: "Only in 18% of these cases all subjects agreed, whereas in 82% of cases subjects disagreed to some extent whether or not a pair of images should be considered near-duplicate".

Nevertheless, advances in data structures, algorithms and machine learning have allowed large technology-driven platforms to shift away from simple content identifiers towards dynamic and data-driven approaches like deterministic fingerprint-based content recognition and automated decision-making systems. For example, in the audio space several content-recognition based

[5] In this paper, we use the term content primarily in an abstract sense. Content is typically an abstract idea (e.g., creative work) independent of its manifestation. When we talk about data or digital content, we refer to specific digital manifestations of content - e.g., bitstreams that can be interpreted and processed by computer systems.
[6] https://en.wikipedia.org/wiki/SHA-2
[7] https://tools.ietf.org/html/rfc2313

technologies have been adopted, including proprietary technologies used by YouTube (Content ID) and Shazam (acquired by Apple).

The shift originates from the insight that the human-curated assignment and management of identifiers based on descriptive metadata does not scale to the amount of content that requires identification.

While content recognition technologies have matured in recent years, they still pose significant challenges and risks. There are numerous and mostly proprietary fingerprinting approaches for various media types with different matching and recognition capabilities. The lack of transparency and interoperability that surrounds these proprietary technologies has caused a competitive imbalance.

5.2.4 International Standard Content Code

5.2.4.1 Overview and summary of use in DisTRi

The International Standard Content Code (ISCC) [18] is a proposal that introduces a new category of identifiers for digital assets that targets most of the DisTRi requirements enumerated above. It combines existing fingerprinting, machine-learning, and shared-ledger technologies to create an open, inclusive and decentralised system of content identifiers and content identification.

In our initial design of the DisTRi Framework, we use the ISCC as a data-driven content identifier that is created from the bits associated with content and (typically) a small amount of initial metadata (e.g., creative work title). The ISCC is then bound to the content's IDB, as described in the first part of this section 5.

However, the full ISCC approach promises much broader functionality and so we describe the approach in greater detail below.

5.2.4.2 ISCC in depth

The ISCC framework is a neutral, transparent, interoperable, and vendor-independent system. All development of the ISCC is organised as an open collaboration effort and all results are available as open-source software with permissive licenses.

Traditional standard identifier systems tend to couple the identity of content with the identity of an actor (any person, corporate entity or group that interacts with the content) requesting an identifier to be issued - for instance by including the identity of the registrant in the identifier code. The ISCC goes beyond this by following the principle of separation of concerns (SoC) [19]. The association between an ISCC and the digital asset does not require a registry and can be generated and verified independently using an open tool - in this case implemented with open-source software. This separation of concerns, which segregates the identity of digital content and the identity of registrants, rightsholders or other interested parties, makes it easier to handle the complex relationships that must be managed in the digital ecosystem.

We assume that in a multi-sided ecosystem, numerous actors may have a legitimate interest to create, lookup or publicly declare an identifier for a given digital asset. Providing personal identity or claiming authorship should not be necessary to acquire a content identifier for a given digital asset. However, having an interoperable identifier is a requirement to communicate and agree on authorship, copyright, and many other possible assertions. Examples of legitimate interests are orphaned content, text authored by journalists in countries with oppressive governments or content that must be referenced reliably even if a standard identifier has not yet been assigned or it is not discoverable.

Just like cryptographic hashes, ISCCs are derived algorithmically from the digital content itself. However, instead of using a single cryptographic hash function to identify data only, we use a variety of algorithms to create a composite identifier that exhibits similarity-preserving properties (like a fingerprint). Furthermore, the component-based structure of the ISCC identifies content at multiple levels of abstraction. Each component is self-describing, modular and can be used separately or in conjunction with others to aid in various content identification tasks.

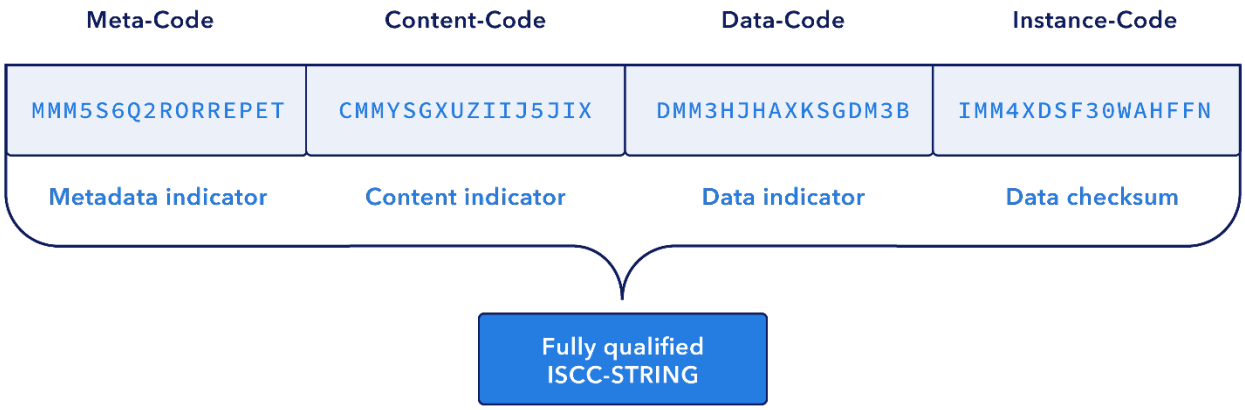

Figure 2: Format of ISCC string

The low barrier of entry to acquire globally unique identifiers for digital content enables interoperability between different actors and systems using digital assets. The algorithmic design supports scenarios that require content deduplication, database synchronisation and indexing, integrity verification, timestamping, versioning, data provenance, similarity clustering, anomaly detection, usage tracking, allocation of royalties, fact-checking and general digital asset management use-cases.

5.2.4.2.1 Levels of digital content identification

In our approach we distinguish between different technologies and multiple levels of digital content identification on a scale from abstract to concrete.

5.2.4.2.1.1 Level 1: Abstract identification

An abstract creation is an intangible idea or concept in the broadest possible sense. The scope of identification is entirely independent of any manifestations of the idea or work. Because abstract

creations are intangible, we can only refer to them indirectly by describing them (reference metadata). An obvious approach is to use a cryptographic hash generated from reference metadata as an algorithmic identifier.

Unfortunately, different actors will describe the same abstract creation in different ways, even if they use a standardized metadata schema. Differences in spelling and punctuation will yield uncorrelated cryptographic hashes. For these cases, the ISCC provides a purpose-built component (Meta-Code) that improves the robustness of the hash-code by using text normalization and a similarity-preserving hash (a function that preserves correlations between similar inputs) (SimHash) [20] generated from the reference metadata. This approach supports the detection of errors and can be used to cluster digital assets based on similar metadata.

5.2.4.2.1.2 Level 2: Semantic identification

A semantic embedding is the machine-readable representation of a higher-level concept or idea. Advancements in machine learning and deep hashing techniques have enabled us to extract and codify abstract concepts manifested by digital text, images, audio, or video.

For example, in recent years, we have seen multiple and increasingly successful attempts to create cross-lingual word and sentence embeddings [21]. Taking these cross-lingual embeddings and using them as inputs to a similarity-preserving hash function, we can construct the same or similar hash-code for a given text even if it has been translated into another language. While this is still an area of active research, we have already carried out tests that suggest the general feasibility of the approach.

5.2.4.2.1.3 Level 3: Perceptual identification

At this level of identification, we are concerned with the syntactic, structural, or perceptual identity of the content. For example, given a PNG and JPG version of an image, the data of these files will be different while their perceptual identity remains mostly the same. Accordingly, the purpose of identification at this level is to gain robustness against a number of non-adversarial transformations regularly applied to digital content (e.g., transcoding, scaling, compression).

Algorithms for codifying perceptual identity must use different strategies depending on the generic media-type (text, images, audio, video) that needs identification.

The ISCC makes use of various granular fingerprinting algorithms (Phash [22], Chromaprint [23], Mpeg7 Signatures [24]). These fingerprints are the inputs to the hash-functions that create the short and content-aware binary codes that we call Content-Codes.

5.2.4.2.1.4 Level 4: Data identification

At this level, we identify digital manifestations of content with a specific encoding and content format. However, unlike a cryptographic hash, the algorithm is tolerant to some variation in the data to allow for minor edits and updates without creating a new or uncorrelated hash-code.

The algorithm to generate ISCC Data-Codes first creates shift-resistant data chunks (FastCDC) [25] from the digital asset. We then hash the data chunks using a fast non-cryptographic hash function (xxHash) [26] and use a similarity estimation technique (MinHash) [27] to create the final Data-Code.

5.2.4.2.1.5 Level 5: Data verification

At this level, we identify the exact binary representation of a digital asset by using a Merkle-tree [28] based cryptographic hash function. It serves as a checksum for the digital asset and can be used to verify the integrity of the data or find exact duplicates.

Depending on security and uniqueness requirements, applications can generate and store Instance-Codes of different lengths. Longer variants of an Instance-Code are extensions of the shorter versions and remain compatible. The use of tree-based hashing supports efficient containment-proofs and verified streaming. A data delivery application can provide cryptographic proofs for partial data and streams such that the receiving application can incrementally verify chunks of data against a given Instance-Code.

5.2.4.2.1.6 Level 6: Individual copy

In the physical world, we would call a specific book (one that you can take out of your shelves) an individual copy. These are termed “items” in both the metadata frameworks noted in section 4. An individual copy implies a notion of locality and ownership. It is also associated with the context of ownership or custodianship: different copies of the same content may be associated with different playlists, annotations, preferences, or attitudinal metadata such as mood or genre. Individual digital copies may be distinguished by their location on a specific physical storage medium. But the concept of ownership gets blurred in the digital realm because data can be replicated at near-zero cost and without affecting the original copy. Accordingly, we might distinguish a digital copy by a license or by some data access policies in the digital realm. However, with the emergence of non-fungible tokens [29], it is now possible to have cryptographically secured, publicly notarized, tamper-proof and transferable certificates of ownership of individual copies of digital content.

6 Rightsholders and Self-Sovereign Identity

The Distributed Trusted Rights Framework (DisTRi) allows for a complex ecosystem of multiple actors and works across different industries and jurisdictions. Around this, a network of interfaces, tools and services can interact with the Framework and support existing businesses or create new ones. We elaborate here on what constitutes an actor, what are the roles that an actor can perform and, in particular, how they can be identified and referenced.

Each work originates with one or more rightsholders who are attributed the rights for authorship or creation. However, even more in the world of digital content than in traditional distribution networks, multiple rights accrue to other actors as the content winds its way through the chain of creation, exploitation, and distribution. These rights vary by creation type and, for example in music, the actors include not only the songwriters but also publishers that may represent them,

collective management organisations that manage the rights (whether by statute, assignment or licence), performers who record the work, record labels who may have a relationship with the performer, studio professionals who may be due payments and other collective management organisations who handle label and performers' rights. The identities of all these actors must be managed.

'Rightsholder' is a role or attribute of an actor who is associated with a work, which to be managed in this Framework is linked to a piece of digital content. More than one such actor may be involved in a single act of creating a work and these actors may have a collective identity (for instance as performer members of a group or band). Within the Framework whenever an actor asserts that they are owner of certain rights, they must have some unique identification, so that the rights can be unambiguously attributed, and eventual remuneration can flow.

While the Framework principally deals with rights created by individuals and groups thereof, it also allows for actions made by individuals on behalf of corporate entities, such as an employee of a corporate music publisher. In this case, the individual and the publisher need to have appropriate identification so that the authority for one to act on behalf of the other can be managed.

The Framework does not prescribe the technology used to manage identity, though to submit content to the Framework, use of some agreed mechanism is required. This can be via a number of existing systems which work with centrally issued identifiers, including OpenID Connect (OIDC)[8] (which is an identity layer over OAuth2.0 authorisation layer) and Security Assertion Mark-up Language (SAML)[9]. Self-Sovereign Identity[10] (SSI), which offers a more decentralised identity model, can also be used. This is delivered by the eSSIF initiative [30] within the European Blockchain Service Infrastructure and by other providers elsewhere [31].

DisTRi allows users to use a Creative Passport[11]. This allows services such as DisTRi to offer a 'sign in with Creative Passport' option, using OIDC, to create a link between that service and the holder's Passport. This does not preclude the use of other authentication or authorisation methods but connecting with the Creative Passport allows a rightsholder reliable association (and hence attribution etc.) with the identifiers that are assigned to works with which they are associated.

The Creative Passport (CP) is a tool for managing digital creative personas, a way of collating and managing existing identifiers as used within the creative industries as well as being able to share, publish and link the data.

The Creative Passport, as a tool for managing creative personas, allows existing industry identifiers to be associated with that persona in much the same way as content within the

[8] OpenID Foundation. "OpenID Connect", https://openid.net/connect/
[9] "Security Assertion Markup Language (SAML)", https://en.wikipedia.org/wiki/Security_Assertion_Markup_Language.
[10] "Self-Sovereign Identity (SSI)". https://en.wikipedia.org/wiki/Self-sovereign_identity
[11] "Creative Passport". https://www.creativepassport.net/

Framework can be associated with legacy content identifiers. Identifiers created by systems such as ISNI[12] , IPI[13], and IPN[14] can be associated with a Creative Passport, creating a bridge to existing workflows.

Designed to work in a way that allows the holder of the Passport as much control over how they use the identifiers and information that is held in the Passport as possible, the Creative Passport works with existing identity, authentication and authorisation systems and methods as well as emerging methods, such as SSI. It can act as a 'wallet' style container, allowing for the creation and use of cryptographic key pairs as well as working with SSI style credentials. Creative Passport also works with a more traditional supporting infrastructure to aid in backing up and securing the availability of the holder's data.

The Creative Passport allows the holder to connect this electronic identity with its container of creative career metadata, with a verified real-world identity, representing a singular person. So, the creative personas can be traced back to an individual, though the disclosure of this link is within the control of the individual, and within legal and regulatory constraints.

Each interaction where data is shared from the Creative Passport with another service is logged by the CP, allowing the holder to review where they have shared points of data.

Whilst the CP is focused on the individual, collaborations are (as mentioned above) a key feature of the creative process. Any 'persona' may be associated with others in the creation of a single work. The CP is designed so that these relationships can be documented, allowing 'group' identifiers to be linked to the individual members of the band, collective or team. It accommodates the passage of time and the shifting of members of these groups and their changing roles.

7 Types of assertions and attestations

There are many types of assertions that are relevant here. Perhaps the simplest or at least the most minimal for our context is: "This content exists". The DisTRi Framework provides 1) a way of clearly expressing the statement where "This content" is well-defined, and 2) a way in which we can immutably place existence in some timescale. The importance of a content identifier is explained above and we can use it concretely and succinctly as the subject of the assertion (and will often be used as objects of assertions). Placing the statement in an immutable ledger provides the timestamp.

The more general assertion is a declarative sentence that will have a subject, predicate, and object, and a submitter or declarer. There may be various modifiers, as well. DisTRi does not require any formal language for assertions, but it does provide a means for declaring assertion types, whereby web forms can lay out a template for an assertion in a human friendly fashion, and values in the form fields can then be transformed into value entries for a data type that has

[12] ISO 27729 International Standard Name Identifier (see: https://isni.org/)
[13] Interested Party Identification (see: https://www.cisac.org/What-We-Do/Information-Services/IPI)
[14] International Performer Number (see: https://www.scapr.org/tools-projects/ipd/)

been registered with the DisTRi assertion type and schema registry where JSON types and XML tags can be defined. DisTRi can support other data description languages, as we expect that in a universally distributed data management system, data virtualization[15] capabilities that include translators will be necessary. As new data providers and services appear, the data descriptors will need to be chosen wisely for ease of interoperability. The most important part of assertion parsing is to determine compliance with policy which is explained below.

For ease of presentation only, we will refer to assertion types as exemplified by JSON types, but we are serious in accommodating other data descriptors and schemas.

There are also different *kinds* of assertions each of which can have various formal data types. Several are listed below:

1. Existence: Generally, a simple statement: "*Submitter* says *this content* exists"
2. Origination/provenance/attribution: *Submitter* says: "*Artist* composed *this content"*
3. Content relationship such as inclusion, version, instantiation: *Submitter* says: "*This* is a recording of this *composition"*
4. Property/attribute/metadata binding: *Submitter* says "*This* ebook has *this* metadata set"
5. Group bindings that can define content as part of subscription series, or artists as members of guilds, etc.
6. Compliance attestation such as statements of compliance with sectoral or jurisdictional content policies
7. PKI binding: *Submitter* says "*This* public key belongs to *this* entity"
8. Revocation: *Submitter* says *"This assertion* is revoked"
9. Update, a statement that revokes one version of an assertion and replaces it with another, preserving precedence

The DisTRi Framework also connects to the global network of personal identity providers and services for Self-Sovereign Identity and identity attributes. Some of these use various authentication protocols native to their service, but others can use the DisTRi distributed ledger system, where identity assertions can be recorded in a ledger. In the former case DisTRi can use data virtualisation and/or the native protocols used by those identity providers, according to local policy.

7.1 Assertions, topics, and policies

One of the goals of DisTRi is to provide authoritative knowledge with assurance that is instantly available across a broad variety of topics. The amount of knowledge that needs to be verified in depth when we want to use it to automate systems is prodigious given the kinds of automated systems that people want to build. No single entity can reasonably be given the responsibility for

[15] Data virtualization: use of a common interface to enable granular data access and governance for a diverse set of data sources

originating, curating, and maintaining that knowledge. Furthermore, the appropriate level of security, and the depth of the verification of entries into the ledger will vary according to the type of information. Therefore, it is necessary to distribute responsibility for policies that can be used to verify both the identity of someone making an assertion as well as their authority. The most straightforward way of doing this is by having different ledgers with different policies, and to support different trusted identity providers. The policies can refer to both domains of authority and knowledge topics. Domains can include Internet domains but go beyond that with additional types of domains and subdomains that can be registered. *Assertion Topics* within domains and standardised topics that are registered across domains can subdivide areas of knowledge. DisTRi therefore has a registry for domains and topics that can be used to register assertion types that belong to different domains. These registries can also record ingestion policies that reference the knowledge that can be expressed in the various assertions. That is, when someone submits a DisTRi assertion, that entry can be determined to correspond to a specific topic based on the prospective assertion's content. The policy for the specialised ledger can then be used to 1) verify the submitter's identity, and 2) parse the submission to extract domain and topic, and then 3) determine whether that submitter has the authority (permissions) necessary to make that submission given the domain and topic. In a mature system, the ledger here will appeal to other ledgers or their derivative databases (see below) to execute the steps above. For example, identity can be validated by checking a binding between an identifier and a public key, or hash of a secret used in a login protocol. The policy could, in some cases, require multi-factor authentication. The ledger could reference another specialised ledger to determine the authority of the submitter.

7.2 Assertions and bindings

There are many other kinds of assertions as well as variations of the ones informally described above. Generally, except for the first, the various kinds of assertions above will be implemented in DisTRi as a binding of data fields where the meaning of the binding is declared in the type registration. But the binding is realised as a cryptographic hash of a serialised concatenation of data fields.

In the DisTRi Framework, a *binding type* describes a simple form of an assertion. It is a description of the components of a binding, along with a description of the meaning of the binding. As with all assertion types, a binding type can be local to a domain, or it can be registered across domains in a global registry. Here is an example:

- Binding type registered name: GlobalCopyright_original_authorship_claim
- Semantics: Author_ID claims original ownership of copyright for Content_ID
- Components:
    - Registered_Author_ID_provider_name
    - AuthorID
    - Registered_content_ID_type
    - Content_ID
    - Additional submission metadata: SubmitterID, Time of submission, etc.

This binding type is easy to describe in JSON, XML, YAML, etc. and an instance is straightforward to serialise and hash.

8 Previous approaches for providing certified, authoritative information

Previous approaches for certifying data authenticity and authority have used digital certificates. And while DisTRi uses digital certificates when appropriate, we favour the use of trusted, shared ledgers[16] with the properties described in subsequent sections. However, it is important to review the technology that is in current use.

Digital certificates are well-known for their role of binding public keys to ID, as is described by the X.509 standard[17]. The most familiar example is the binding of a Uniform Resource Name (URN) to a public key that certifies the key used in the Transport Layer Security (TLS) protocol used for the HTTPS protocol. The certificates are typically issued by a Certificate Authority or CA, and applications that rely on the binding use a sequence of keys that can be traced to the certificate authority "root of trust". Less well-known, but just as effective in many ways is the use of SAML (Security Assertion Markup Language)[18] and XACML (eXtensible Access Control Markup Language)[19] certificates to bind other types of information allowing validation of attribute assertions.

X.509 certificates are portable, and we have thirty years of experience working with them. They are issued by certificate authorities and typically have a specified period of validity. But they are hard to renew and revoke, and do not provide a capability called “Perfect Forward Authentication”, since a compromise of a key used to sign a certificate invalidates all documents signed after the compromise but also before the compromise, and it invalidates all certificates in the key chain below that key both before and after the compromise. So, generally we would only want to use certificates where trust is required for a relatively short period of time, thus the importance of an expiration date. However, expiring certificates have often caused a lot of grief (for reasons we will not cover here), and it is common to ignore the expiration date. Loss or compromise of high-level or root keys can cause major disruptions, and so for serious CAs, it is necessary to have extensive security measures defending against compromise and provisions for disaster recovery, as well as techniques like using "intermediate keys" for signing that, while effective, add to the complexity and computation load of systems that use certificates from such CAs.

See [32], [33], and [34] for examples of how the complexity of key management using these kinds of certificates can cause pain and disruption, even in relatively small Public Key Infrastructures

[16] Shared ledger: agreed-upon, time-ordered, append-only sequence of signed data replicated across multiple computer systems operated by independent actors
[17] https://en.wikipedia.org/wiki/X.509
[18] “Security Assertion Markup Language (SAML)”. https://en.wikipedia.org/wiki/Security_Assertion_Markup_Language.
[19] “eXtensible Access Control Markup Language (XACML)”. https://en.wikipedia.org/wiki/XACML

(PKIs). In the case of web browsing applications, we have been able to deal with the pain so far. However, business and industrial automation and an authoritative web of copyright information will need to be much more resilient and scalable. In all these years of using X.509 certificates, we have not yet seen anyone deploy an efficient system for renewing and revoking certificates in the commercial domain. With strict information hierarchies that mimic the hierarchical nature of X.509 design, it has been done, but in open commercial applications, it has been very difficult. The two main approaches to revocation are CRLs (Certificate Revocation Lists)[20] and OCSP (Online Certificate Status Protocol) [35]. The latter requires checking a certificate and every certificate up a certificate chain to its root with an online authority. Recent advances (including a technique called Online Certificate Status Protocol (OCSP) stapling [36]) have made this approach more scalable. Nonetheless, for DisTRi we are talking about an infrastructure that will eventually be thousands of times larger in certified data items and for each content item we may need to certify dozens more attributes. So small improvements in scalability will not work. All that said, X.509 certificates will have their role in the future and in DisTRi.

9 Distributed systems; trusted, distributed data management

The term "Distributed System" is overloaded in computer science. In this paper, we will refer to multiple planes of distribution:

1. Services categorised by specialised functions, some supporting or provided by different DisTRi actors, such as the producers of content, the users of content, or providers of identity, credentials, metadata, etc.
2. Services categorised by content type or commercial sector: music, video, news, photography, design information, etc.
3. Producer or consumer orientation: Some services are oriented to vet and accept data submissions, and others that group and organise information from multiple sources and organise it for efficient discovery and use by information users
4. Redundancy: Multiple copies of information stores and policy-compliant processing provided for the benefit of resistance to error, malicious or benign
5. Jurisdiction: Services that differ by the customs or requirements of different jurisdictions

The major challenge that DisTRi addresses is to allow many thousands of different services to support and interoperate with one another through standardised, yet extensible service APIs. As suggested above, this is practically required to accommodate the many different kinds of specialised knowledge and authority that must be referenced in order to support the aims of a global copyright infrastructure.

A second major challenge for DisTRi is to govern the information that is submitted to the system. DisTRi could be thought of as a cooperative of stakeholders, some of whom supply information, some of whom use it, but all of whom have major stakes in assuring that the information is

[20] "Certificate revocation list (CRL)". https://en.wikipedia.org/wiki/Certificate_revocation_list.

properly protected and used only according to agreed-upon or advertised policies. A trusted, multi-party, distributed data management system is then required to:

- Securely manage data end-to-end, ensuring that it does not escape from its sphere of governance
- Allow data to reside in different sovereign physical locations while serving multiple stakeholders with divergent interests
- Enable computation on multiple data sets without the need to move that data from its native physical locations
- Provide comprehensive and precise Identity and Access Management (IAM) capabilities covering people, devices, and software services such as analytics, establishing permissions for various actions that depend on explicitly authorised attributes
- Scalably accommodate dynamic data from multiple sources; this is required to meter content usage
- Maintain the provenance and authenticity of data, end-to-end, as the integrity of content and content metadata requires assurance of data provenance
- Ensure that attributes and other claims about data and connections to entities (such as ownership, and rights) are authoritative
- Provide the ability for additional parties to operate on the data with specialised applications, producing new knowledge governed by policy and precision IAM. This is required for many of the applications that will use DisTRi applications. This includes auditing and analytics that may in some cases be private and in other cases public.

The requirements above can be implemented by assembling several technologies and capabilities from data management systems providers. However, the DisTRi project has begun an *initial* implementation based on the Intertrust Platform[21] which can be described as a trusted multi-party distributed dynamic data management system that features methods and capabilities that can be used to build the kind of system components that DisTRi uses to meet the objectives and requirements described in this paper. In particular, this Platform provides an integrated set of capabilities featuring:

- *Governed Data Virtualization* which is essential to implement a system that integrates many thousands of independently operated services and specialised areas of knowledge and confidential data
- *Precision IAM and Data Governance* that can apply fine grained policy to requests for action to create, read, update, and delete assertions, and other specific data items in a globally distributed system
- *Secure Execution Environments* that can support controlled interactive data exploration and processing while providing privacy and confidentiality protection; these

[21] https://go.intertrust.com/hubfs/assets/Platform/Intertrust-Platform-Quick-Guide.pdf

environments providing support for third party processing of sensitive data without revealing the raw data

- *Scalable Time Series Database Processing* that can combine the above capabilities and implement such capabilities as meters that measure content use and produce detailed and governed data records for stakeholders in content distribution
- *Robust and detailed logging and auditing tools.*

The platform is used for multiple DisTRi functions, but specifically for the implementation of the DisTRi ledgers that accept assertions, metadata, identity information, etc. while providing all the properties of authenticity, accurate provenance, and authority. Thus, the major components of the DisTRi Framework are what we call TIDALS and TIDAL derivatives. These are described next.

9.1 TIDALs

TIDALs provide the basic capability for recording and retrieving various assertions that comprise the Copyright Infrastructure (CI). The term *TIDALs is an acronym: Trusted, Immutable, Distributed, Assertion Ledger System*. The terms are illuminated below:

- AL: An *Assertion Ledger* is a database of assertions. Generally, but not necessarily, these are in the form of fixed value predicate statements as described above, but they can also include bindings with implicit predicate statements.
- D: The term *Distributed* can mean different things. *Distributed Ledgers* has come to mean an implementation of ledgers that have multiple nodes each of which has an identical copy of the ledger content. While the ledgers in a TIDAL System can be and usually are distributed in this sense, we believe that in order to attain the kind of scalability required for the CI and other applications that want to store authoritative, reliable information, the content itself needs to be distributed among various specialised databases where policies associated with access can be specialised to different kinds of content, such as identities for different kinds of actors, and different kinds of content and metadata. We envision thousands of ledgers worldwide to cover the gamut of all types of specialised assertions across various jurisdictions for just one content type such as music. Many more ledgers will be specialized for film, video, text, graphics, etc. and even more for non-media IP such as pharmaceuticals, 3D printing, and the like. So, a TIDAL System is multi-distributed with knowledge and authority distributed across many ledgers, each ledger having multiple nodes, and each node having separate content and verification structures.
- I: Entries into a TIDAL are *Immutable* and cannot be directly revoked. They will be time-stamped, sometimes in multiple ways, and as discussed below, there will be hash indexes that will make it easy to find revoked and updated entries about a subject. But a complete historical record will always be available.
- T: Some distributed ledgers are said to be trustless. We do not agree that any ledger is trustless, even though it might be true that trust is favourably rearranged in some ways. The word *Trusted* here denotes the concept of relying on adherence to policy for a given

ledger. In TIDALs there are at least two policies: 1) the ingestion policy for accepting entries, and 2) the agreement policy or consensus mechanism where multiple parties agree on an entry into the ledger or multiple parties compute the entry using a multiparty computation protocol. We expect that each stakeholder or stakeholder group will run one or more nodes, or delegate an entity to run that node, and consensus will require affirmation or Multi-Party Computation (MPC) participation from each of the stakeholder groups. Consensus in most cases should not require huge numbers of nodes.

- S: The term System used here denotes the fact that we will be using a system of ledgers that are specialised for the policies they use and the services they provide. Each TIDAL is supervised and organised by specialists, or they may be ledgers that are automatically created by an existing authority, including Performance Rights Organisations (PROs) and CMOs. For example, there are a number of existing identity management specialists for different kinds of artists. A simple script can create an initial TIDAL using a simple ingestion policy and an agreement policy that can be updated over time, strengthening its security. This TIDAL can then interact with other TIDALs that require the use of the identities verified through this first TIDAL.

9.2 TIDAL derivatives

In our Framework, we expect to have numerous TIDALs each of which may specialise in the recording of various types of assertions with policy appropriate for those types. We expect that a given application that needs to rely on the authenticity of multiple assertions will either directly or indirectly query multiple TIDALs. We mention indirectly here because our Framework includes the notion of trusted TIDAL derivatives which are databases of information derived from one or more TIDALs. The derivative will typically be focused on a class of applications that have requirements for efficiently authenticating specific affirmations.

It is useful to consider an example here. Suppose there is a TIDAL that specialises in recording affirmations linking IPv6 addresses and domain names (DNS server). A TIDAL entry will include among other fields, a hash

H1= h(IPv6_address | domain_name)

Suppose also that we have a TIDAL that specialises in recording affirmations linking domain names and thumbprints (hashes) of public (cryptographic) keys (see below) for the devices using those domain names. This TIDAL will have a hash

H2 = h(h(public key)| domain_name)

Let's also suppose that we have a TIDAL derivative that is designed to help applications that are dependent on knowing the proper IPv6 address and the public key of a device. Such a derivative could include a component that uses agents to monitor multiple copies of each of the two TIDALs mentioned above. This derivative could maintain a hash table of all hashes of the form

H3 = h(H1|H2)

Such a hash table for a billion domains would, using large (512 bit) hashes, be relatively small. When an application wants to send a secure message to a device with a given domain name, it can compute H3 locally, and look it up in the hash table of the derivative. If it is not found, then the information is not deemed valid, according to policy. Since this is generally a rare case, a secondary search, perhaps to the original TIDALs can reveal the reason for the error (such as a revocation). However, in most cases, the hash is found very quickly (finding even a 512-bit hash in a hash table of 512-bit sparsely distributed numbers (as hashes will be), is extremely fast. This will be much faster than locating and verifying the certificate chains of corresponding Domain Name System Security Extensions (DNSSEC) and TLS certificate entries. Since the hash tables are small, we can afford to maintain many copies, and agents can cross check them, and different policies can determine, for different applications, whether to perform redundant lookups, cache results, etc.

At this point, it is useful to note that TIDAL derivatives can be used to help with an issue regarding immutable ledgers, namely that information can change, and affirmations recorded in ledgers may need to be revoked. This is also a significant issue with other forms of authentication such as those that use digital certificates. When an affirmation is revoked, that revocation affirmation can be included in the same TIDAL as the original affirmation. TIDAL derivatives can include components that monitor TIDALs for revocations and updates. These derivatives can then remove the corresponding hashes from the hash tables they maintain and provide secondary lists that are searched when search errors occur in the main hash table as a result of a revocation.

In our Framework, TIDAL derivatives will function in a different computational context than the publicly accessible TIDALs and will be secured appropriately for the given applications and clientele that they support. They can also be redundant, and we can have multiple, independently maintained cross checking derivatives for any given application. Thus, TIDAL derivatives will organise authentication information, cross check it with multiple copies of each TIDAL it monitors and possibly with other TIDALs that have the same authentication information, and maintain special hash tables for common authentication queries making verification of information authenticity extremely efficient even for affirmations about trillions of entities.

9.3 Leverage and resilience in the TIDAL system

The immutability of TIDAL assertions is a strong property that cannot easily be duplicated by other approaches such as one that relies entirely on digital signatures and certificates, as described above. Once a digital certificate signing key is compromised, all certificates signed with that key, including those signed before the compromise, must be deemed invalid. However, if an assertion is submitted to a TIDAL, even if the assertion was presented by an authority using a digital certificate whose signing key was subsequently compromised, the immutability of the TIDAL will still stand, and fewer assertions will therefore need to be revoked. Most importantly, the TIDAL system, even when it uses PKI to authorise assertions, does not rely on the long-term security of signing keys and their hierarchies.

The use of different ingestion and agreement policies for different TIDALs allows us to apply a powerful principle of security more deftly in DisTRi, namely security *leverage*. Leverage is defined

as the quantity L=(R*C)/B where B is the expected benefit due to an entity that launches a successful attack on the system, C is the cost necessary to launch the attack, and R is the risk (and consequence) of being discovered. A system of specialised TIDALs provides the opportunity to design the appropriate leverage against attack at different points in the system. We can increase leverage by decreasing Benefit, or by increasing Risk and/or Cost. Leverage is often difficult to estimate initially, but another advantage of using the TIDALs approach is that we can adjust these parameters over time. TIDALs can over time accept increasingly more evidence when an affirmation is submitted, including multiple certificates, and many other kinds of evidence including proof of location, coercion-evading protocols, etc. They can increase the number of entities necessary for agreement so that attackers will need to subvert more of those entities, stronger protocols can be implied in each case without requiring major changes in the rest of the system.

TIDALs that use blockchains and other data structures for immutability can employ a meshing strategy that increases resilience to attack. While TIDALs in DisTRi are independently operated, and some TIDALs may involve only a small number of operating participants and redundant nodes, they can cooperate by accepting blocks from one another on a periodic basis. DisTRi encourages TIDALs to cooperate in this manner, providing cheap but effective error resistance. Some TIDALs may even find this strategy useful in fully replicating their blockchains beyond the native nodes operated by stakeholders. None of this need (or should) be done on a single assertion basis, so the validity and authority of assertions still depends on a robust assertion ingestion policy and implementation.

The use of TIDAL derivatives provides high efficiency from massive precomputation of current hash indexes that can be used by applications for rapid validation of data. However, this introduces the possibility of single points of failure unless derivative services are periodically checked, and the services themselves provide for some redundancy. See discussion of the oracle problem in [37]. However, we note that TIDALs are not by themselves oracles with single points of failure under the definitions there, but unless redundancy and independent, real-time testing strategies are employed with derivatives, the latter could easily have such failures.

We emphasize that both the meshing strategies mentioned above, and the automated testing strategies can be deployed, limiting expanse and frequency according to the principle of leverage.

10 Markets, decentralised trust and policies

As explained in the previous section, DisTRi is designed to be highly distributed in a number of ways, and the main driver for this is the need for specialised policies for different kinds of knowledge, different actors, and different kinds of stakeholders. A single TIDAL distinguishes itself by the following:

1. The assertion types it accepts
2. The ingestion policy it enforces for accepting an assertion
3. The agreement policy for recording the assertion
4. The read access policy for ledger entries
5. The data structure used for time stamping and data authentication

Since our Framework allows many independently operated ledgers with localised ingestion and agreement policies, the ledger operators can apply security and integrity principles such as KYC or "Know Your Customer", and attack surface minimality. They can, over time, increase security for their ledgers without severely affecting interoperability. New TIDALs can arise, as needed, driven by market forces, and they can employ meshing techniques to strengthen one another. The DisTRi Framework does not specify any of the policies mentioned above but provides ways in which different ledgers can refer to one another both despite and because of the different policies. Figure 3 below shows where the two types of policies operate in the ingestion of new assertions to a TIDAL:

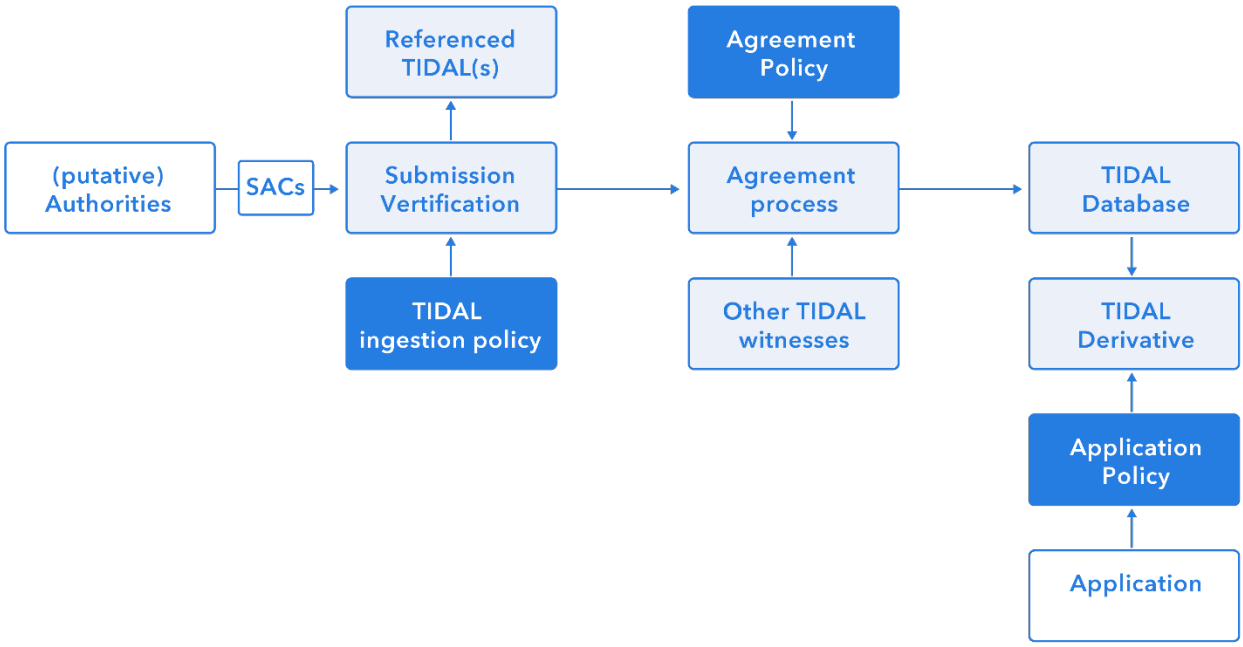

Figure 3: Processing of trusted assertions in a TIDAL

Application policies

The domain/topic approach to organising information seems inherently hierarchical, however in automated processes that check for validity, we may need to reference a non-hierarchical set of assertions. In one application we are working on, where we want to attack the problem of "fake news", we know that videos go through a complex "chain of handling and control". Relevant principals will include the video sensor maker, media module maker, camera maker, internal software processing modules, external post-processing modules, distributors, compliance organisations, etc. We will need to check credentials of processors, authority and reputation of operators, assertions about the robustness of software modules, etc. These will involve several different organisational hierarchies. So, looking at this from the verification point of view, assertions will not be hierarchical, but applications will want to have a direct and trusted source of authoritative knowledge to obtain validation of information and satisfaction of policy necessary to authorise actions. Operators of TIDAL derivatives provide that close and trusted relationship, and they can effectively pre-process common validation queries.

With DisTRi, TIDAL derivatives can seek to organise and pre-validate information from many ledgers, and an application need only rely on one source for this information, but in general applications can use as many TIDALs as necessary, even redundantly, as depicted in Figure 4.

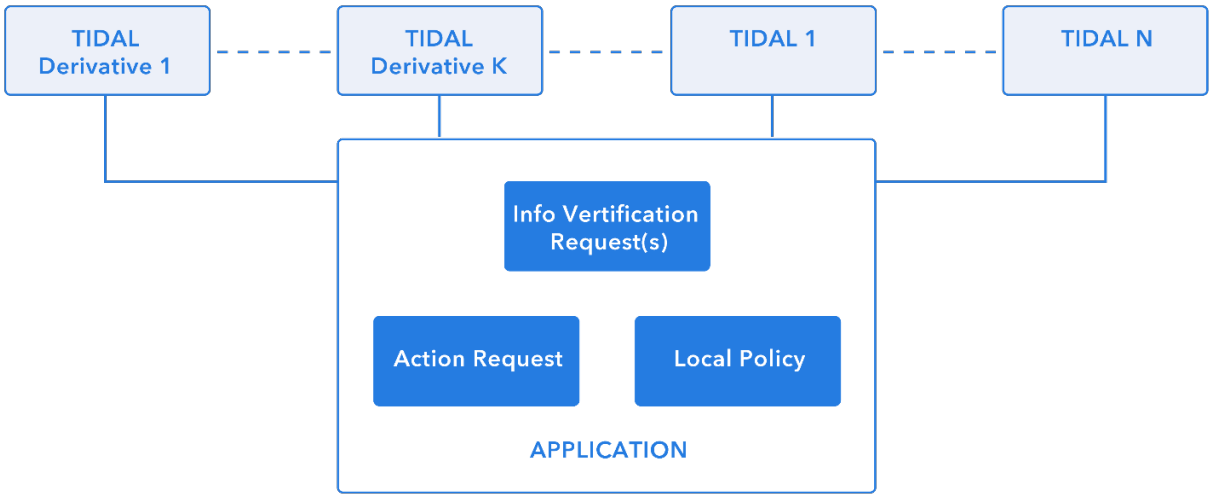

Figure 4: Application seeking authorisation for action

Indeed, a TIDAL derivative will generally be an application that sources authenticated information from multiple authoritative sources, including TIDALs and other derivatives. So, the large box in the diagram above could be, among other things a derivative. Figure 5 below illustrates how the ledgers in DisTRi will evolve, where knowledge, policy, and applications are highly distributed.

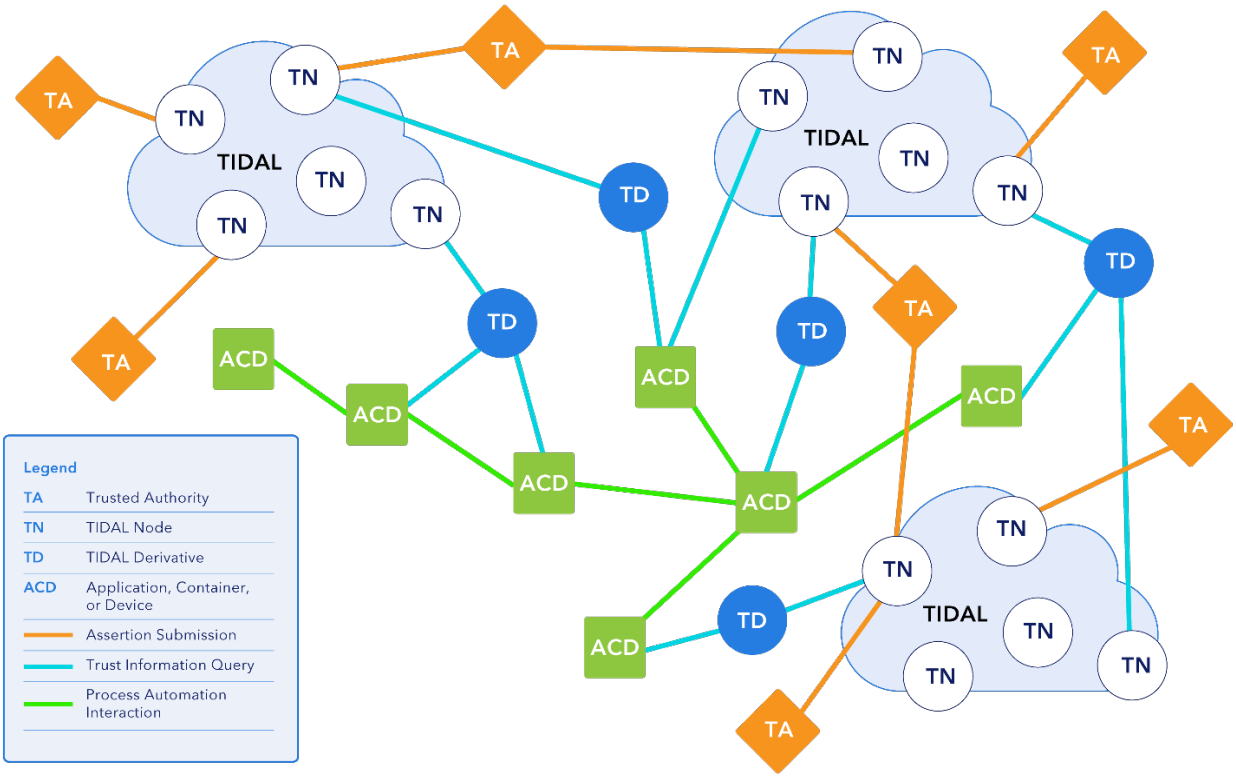

Figure 5: Global trust infrastructure with many specialised TIDALs

11 Human-centric approach

One of our key goals is to allow users of our system, who may be creators rather than rights administration specialists, to accomplish the tasks at hand with a minimum of training or study. If a songwriter, journalist, or other creator has made something, there should be as little friction as possible between them and their audience and between them and their remuneration.

Whether a creator has written a song or a news report, what do they want or need to do?

- Protect authorship so that no one else can claim it
- Reach an intended audience with as little friction as possible
- Be paid fairly for the work.

How will DisTRi enable this for all authors, not just those with existing representatives and legal arrangements in place?

First, we make it easy to deposit the work into an online storage facility where the creator themself can access it at will. As soon as they create it, they can immediately protect it. This deposit should make the attestation that they have done this at a specific date and time, and it can be proven in the event someone wishes to contest that fact – e.g., in cases of copyright dispute.

Further, when they wish to share this work, it is important that all the information stays with the work all the way down the distribution chain. This should be done to enable both credit and remuneration. So, the first job of the DisTRi system is to enable this functionality.

However, the Framework is worthless if creatives do not use it. So, the whole Framework must be intuitive from end to end. Areas of interface:

1. If a creator wants to be paid, they will likely be willing to put up with a bit of work to establish their identity. Currently, creating an identity and using it for ongoing transactions is pretty mature. As discussed elsewhere in this paper, Creative Passport makes this straightforward and Self-Sovereign Identity is maturing quickly. Though there is always a trade-off between security and convenience, no new work is needed in this domain.

2. The creator wants ease of storing and accessing their creations. If they have a great idea while walking through the park, they should be able to record it on a phone and deposit it into the cloud with a couple of taps. This action alone should serve as proof of creation. It can be signed and hashed but this should be invisible to the creator who should just know that it is safe and has been documented

3. The distribution and remuneration systems for different classes of creation are disparate and complex. Payments to creators sometimes come direct from users, from commercial entities with which the creator has a relationship (such as publishers) or from collective management organisations that manage rights and payments where it is too complex or expensive to do so individually (such as photocopying royalties or payments for broadcasting rights). Sometimes the same activity may be managed differently in

different circumstances because of some nuance in the mode of exploitation or differences in the territory of use. Sometimes a creator will establish a relationship with a publisher that is limited in some way (by genre perhaps) and other creations bypass that deal. All these options need to be accommodated and the appropriate memberships, registrations and deposits made automatically. Crucially, the metadata associated with the creation must remain linked to the creation so it can be used whenever it is needed.

Regular users will insist that this complexity is hidden from them but, if someone wants to know how that system works and what it is doing, that should be transparent for them. For the most part, the underpinnings should not intrude on the creative process or the distribution process.

4. Existing identifiers must be supported. While our Framework uses new kinds of identifiers, we must also support legacy identifiers. Many existing systems use these legacy identifiers, and they should be available for use. We support this by binding legacy identifiers with the new identifiers to accommodate existing use cases in new ways and new, yet to be conceived, use cases while not disenfranchising existing participants in the value chain.

5. Both new and existing workflows must be supported. There are many ways that creators work together and communicate with their business partners. The Framework must be flexible enough to support older mechanisms like email as well as newer approaches like creating together on collaboration platforms with numerous people contributing at the same or different times. The Framework must support changes in the way works are created over time and accommodate new and unanticipated members of the value chain. it must also support an unbounded variety of distribution and remuneration mechanisms as they emerge.

Having said all the above, there is also a need for the human-centric approach to apply to system administrators and developers. Consequently, our Framework must have well designed and documented APIs, open-source implementations in multiple programming languages, catering to developers with best practices of modern software development.

Finally, all users of the system must have confidence that their tasks are being carried out with security and robustness so that they do not have anything to worry about.

12 Position in the media and content ecosystem

DisTRi, the Distributed Trusted Rights Framework is the core component of an open, federated, decentralised, and transparent copyright infrastructure.

Users will declare rights into DisTRi and receive attestations in return. Numerous not-for-profit or commercial, sectoral or generic, national or international competing systems will help creators, rightsholders or their appointed intermediaries declare rights related to works and other protected subject matter.

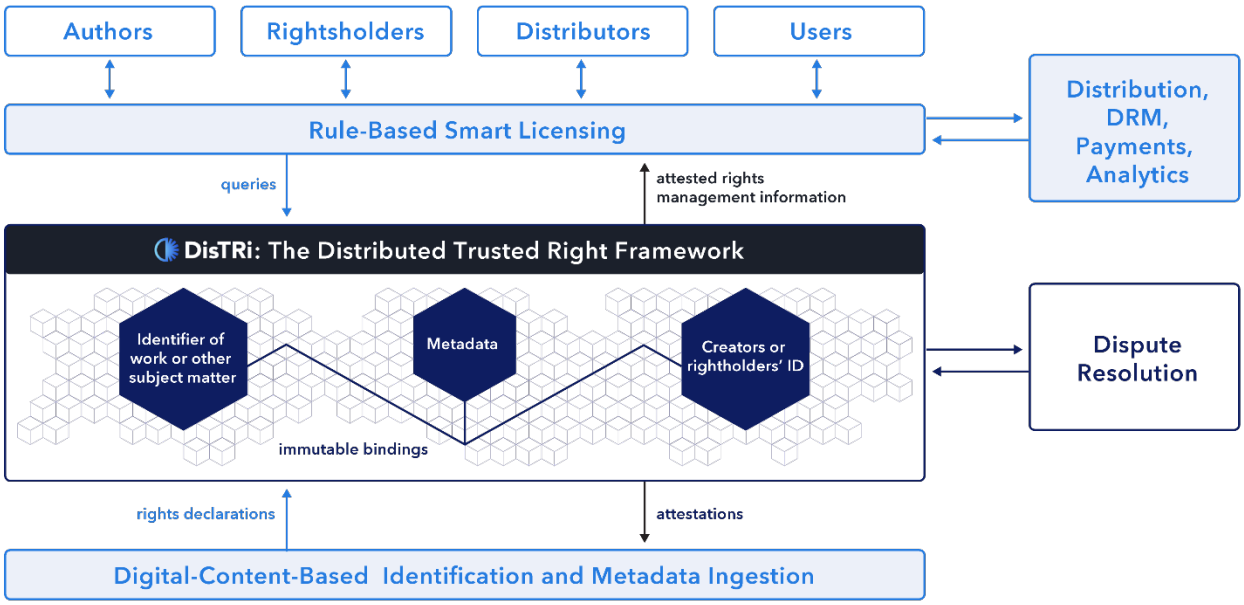

Figure 6: Position of DisTRi in the media and content ecosystem

These works can be new creations. In this case, Rights Management Information (RMI) is current and can be captured in real time at the moment of creation. But the works can also be catalogued productions. Then, RMI must always be checked for accuracy and completeness, even more so if these works are parts of cultural heritage, out of commerce recordings, or orphan content.

Historic rights metadata are often missing or erroneous and have been stored on any medium according to any taxonomy. Therefore, considering the necessity of swift and affordable declaration of rights, humans ought to be aided by machines to prepare RMI for correct ingestion into DisTRi. This computer-aided curation of rights metadata will generate valuable by-products: clean metadata ready to be enriched for indexing or discovery purposes – see Section 4 – or specific sets of metadata for one distribution platform or the other.

The DisTRi Framework will also be useful to solve disputes. Although automated dispute resolution can only be a long-term objective subject to adaptations of sectoral, legal, and judicial practices, DisTRi can already flag anomalies – conflicting claims of ownership or under or oversubscribed attributions. DisTRi can flag the issue and notify the concerned parties inviting them to solve the eventual dispute.

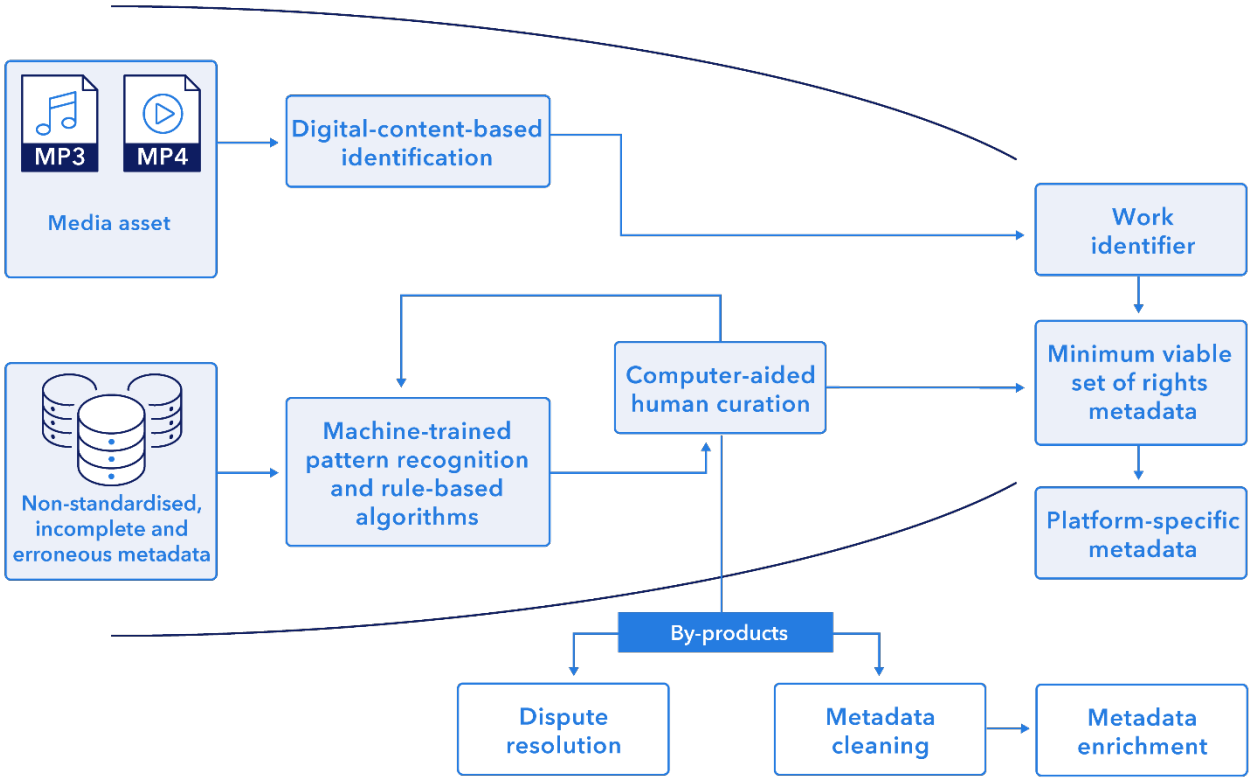

Figure 7: Computer-aided curation of rights metadata

On the other side of DisTRi, users will query the Framework and receive attested RMI in return. A multitude of competing systems will help them license and distribute media assets, monitor and meter their usage, and pay fair royalties. Similarly, these systems will be focused on a sector or cross-sectoral, focused on a jurisdiction or open to the world. Some will be open, some will be proprietary, some may even be internal. This is the domain of Electronic Markets for Media Assets – online marketplaces for ready-made products or creative ideas. Licensing will be automated step by step. For a start, discovering where an adequate media product is located or who could be the appropriate distributor will already be a progress. Then, producers and distributors will be able to rely on configurable contractual templates to generate licensing agreements. Finally, automated contracts will trigger the emergence of new promising business models such as micro-licensing which could not be economically managed by humans.

13 Examples of extensibility and market sensitivity

Here we discuss a few examples of how DisTRi can be extended by others to address some important aspects of content management. One could legitimately argue that these and other capabilities might be more central parts of a framework like DisTRi, however we believe that markets can at least in this case do a better job of determining how to approach some of these applications, either by intrinsically extending DisTRi or by building applications that use resources provided using the Framework.

13.1 Custom meters

Use of digital content is often metered in *closed* distribution systems for numerous purposes, including direct remuneration of rightsholders, computation of proxy payment amounts (advertisements), assessments of effectiveness of distribution, and many others. Private, closed platform streaming services for entertainment content access large catalogues of content and use general agreements, rather than individual contracts to remunerate artists for their work. These services also receive money through advertising and subscribers. Actual playback events affect the pay-out to artists, but the processes for remuneration are thought to be flawed and limited by the availability of reliable and precise information [38] [39] [40] [41] [42], among other things. This is a complicated and even emotion-ridden topic that we do not address here. However, we want to discuss how DisTRi can support applications that can collect event information in both public and private environments, count and classify events according to criteria that reference DisTRi registered metadata types, and govern the dissemination of results, preserving auditability, privacy and confidentiality.

DisTRi does not decide any of the criteria or any of the access policies. It only provides the means whereby these criteria can be reliably referenced, and distribution of meter values can be governed. Third party services, therefore, can be designed to allow for custom meters that use the trusted multiparty data management platforms described above in section 9. These platforms can receive raw event data from multiple sources, and reference both publicly and privately available identity and metadata information stored in TIDALs operated by multiple organisations. Custom, but auditable meters can then be created using that metadata. Public and private auditability of events can rely on automated assertions indexed using hash tables maintained in a ledger derivative application. Hash table entries can provide a common index to different governed views of the same event data, hiding or exposing different information, but allowing reconciliation by both automated and human auditing agents. That is, DisTRi supports ways whereby different stakeholders can reference and discuss the same object, while having faith in the authenticity of the object, preserving privacy and confidentiality.

13.2 Automated contracts

DisTRi does not provide direct means for automated contracts [43]. The minimal sorts of assertion ledgers that DisTRi uses are not designed to include contracts, and DisTRi is not transaction oriented. Business logic is another whole realm requiring specialised trust to administer. However, DisTRi does support automated contracts used in applications that use TIDAL derivatives. This is because a TIDAL, providing signed assertions that support contract conditions, can have customised ingestion and agreement policies that mitigate risks that such conditions may be in error. In fact, DisTRi does allow the use of all sorts of blockchains and other distributed ledgers including those built on Ethereum[22], HashNET[23] and Hyperledger[24] among others. But DisTRi supports business automation architectures that rely on precise but specialised

[22] https://ethereum.org/en/
[23] https://www.tolar.io/hashnet/
[24] https://www.hyperledger.org

trust graphs, where the risks of specific applications can be analysed and addressed. Future papers and demonstrations of DisTRi will show how to design business automation using the authenticated, authoritative data that DisTRi provides.

13.3 Detecting deep fakes

Provenance of data is essential for understanding its trustworthiness, which in turn is important for many purposes such as training AI algorithms, for estimating insurance risks, and for identifying deep fakes. Since data is so malleable, we also need to know how data traverses a chain of handling, how it is transformed, and who transforms it. To accomplish that, we need an infrastructure enabled by a framework like DisTRi that can provide assurances about data. This infrastructure will need to scale massively, given the huge amount of data whose provenance and transformation we want to understand, including image and video data.

DisTRi provides strong and efficient support for applications designed to detect deep fakes by ascertaining content provenance and tracking the chain of handling and control of content objects. As explained in [44], one can use scalable assertion ledgers at every stage of content transformation to ascertain how the content was transformed with permission and legitimate purpose, so that we can flag a version of the same content that is altered without permission or without legitimate purpose. TIDAL assertions can be used to validate the original video sensor, sensor module (with integrated GPS and time base), integrated device (camera, mobile phone, surveillance device), internal post-processing methods (filtering, compression, etc.), and external post-processing methods (specific photoshop operations including cropping parameters, etc.). Professional photographers and videographers, news organisations, and systems and methods designed to provide forensically robust information (body cameras, inspections, etc.) can use DisTRi to document both provenance and chain of handling and control. Broadly applied to many different kinds of sensors, this will produce prodigious amounts of information, and therefore it is important to make sure that the ledger system for storing and referencing the content identifiers and metadata bindings scales well.

Experiments at Intertrust Technologies have shown that this approach using similar components to those in DisTRi scales even for the large volume of videos updated daily to Internet services such as YouTube. In those experiments, videos were divided into short segments (a second or so long), and the content was hashed with compressed forms of the metadata, and each segment hash was entered into a TIDAL, and a derivative hash table stored those hashes. A standard browser with a plug-in could then find the metadata for the video, compute the hashes, and compare the resultant hashes in real time with the hash table service. Since the latter validating operation at the derivative server only requires a small constant time hash table search, per video segment, and the hash is small compared to the data load, and the metadata information is compressed (it rarely varies for different segments of the same video), this system should scale well. Again, DisTRi does not provide this type of service, but the capabilities that DisTRi does provide, and which are used for many other applications, will also support this type of system.

DisTRi's support for the use of similarity hashes in content identifiers, explained above, is also important for designing systems that detect and evaluate deep fakes and other unauthorised

content alterations, whether or not subject content is distributed with documented metadata chains as explained above. While we have not yet experimented with an approach using such methods, an ISCC identifier computation can be used by web crawlers for various content types, and the results can be compared to DisTRi entries, and flagged when comparison scores exceed a threshold. Metadata in registered versions of the content can then be used for further investigations. This approach can be used for applications that detect copyright abuse that is far broader than what has become known as deep fakes. Granular fingerprinting, which is a requirement for more detailed and partial content matching, is also on the roadmap of the ISCC. Interoperable, non-proprietary fingerprints for all media types will enable the development of new and inclusive content recognition systems. In conjunction with DisTRi, these systems will be capable of answering questions like: "What is the earliest known and most trusted (original) incarnation of a potentially manipulated media asset?"

13.4 Interoperability

It has been noted throughout that DisTRi is built wherever possible on existing open standards, albeit configured in novel ways. The support for legacy content identifiers ensures maximal interoperability with systems using these widely deployed standards and the ISCC is an open specification that is being considered for standardisation in ISO. Creative Passport builds on OpenID Connect and OAuth 2.0 and accommodates co-reference with existing name identifiers such as ISNI, IPI and IPN.

But this is not itself enough to guarantee interoperability. Figure 1 sets out a high-level view of the DisTRi architecture and shows a “web interface” between prospective users and the elements that make up DisTRi. For some such users this will, in accordance with the human centric approach above, be a web-browser interface, responsive and tailored to their day-to-day needs. But for others, this will be an API allowing vendors of specialised software and services to access the DisTRi capabilities on the same terms as browser users but at scale or otherwise customised for their user base. These APIs will be open and enable the creation of multiple, granular and specialised services with close attention to the needs and preferences of different sectors.

In establishing these APIs, DisTRi will look closely at existing industry standards to reuse that which has been proved and stress tested. If text publishing standards have, in ONIX[25] created interfaces and standards that can serve the needs of DisTRi, there is no need to reinvent them - indeed to do so will deter adoption. Similarly, the standards of DDEX[26] may serve commercial music applications. Otherwise, new APIs will be needed and “buy-in” from the affected sectors will be sought to maximise utility.

Turning to Figure 6, the functional (as opposed to sectoral) classification of the interfaces to DisTRi shows activity related to rights declarations, attestations, queries and attested rights management information. Where these can be re-used from existing specifications, they will be, but others will be built anew. Where these new interfaces are critical for the operation of DisTRi

[25] https://www.editeur.org/8/ONIX/
[26] https://ddex.net/

or where they have applications elsewhere, they will be considered for promotion as candidate international standards, compliance with which will be straightforward to judge.

14 Conclusion

The task of assembling and distributing Rights Management Information for all the digital content that can be created or digitised worldwide ought to be something that modern Web technologies and cloud computing would easily solve. Indeed, these technologies have been highly effective at making all sorts of information generally available, and at least for many applications, well-organised.

However, the Web also contains a lot of misleading, inauthentic, out-of-date, and overall untrustworthy information. Attempts at finding, organising, and applying information that may come from millions of different sources can produce more noise than signal. Worse, using such information without understanding its provenance, authenticity and authority is often counterproductive and potentially detrimental.

This paper has presented ways in which we can better recognise and organise relevant information and knowledge from millions of sources that is authentic, authoritative, timely, and actionable.

We have shown that the DisTRi Framework can facilitate the assembly of large numbers of resources that can become increasingly trustworthy and responsive to market needs for all kinds of information necessary for media and other data markets to function reliably and efficiently.

We have sought to be minimally prescriptive but maximally supportive and inclusive by allowing many solutions to be employed while enabling many ways in which individuals and organisations can cooperate in originating, enriching, governing, and distributing trusted information.

Our approach has been to assemble new, breakthrough technologies that focus on the fundamentals of content rights management, namely content identifiers, stakeholder identifiers, content metadata associations, authoritative rights assertions, and the use of many specialised, trusted, multi-party distributed dynamic data management systems that can be configured and deployed by anyone who can provide useful capabilities. Rather than featuring a single "distributed" ledger as the centrepiece of this data management system, we allow for thousands of specialised, independently managed but interoperable ledger systems that can each be originated for and adapted to different jurisdictions, creative sectors, or knowledge specialties.

Starting with new, open content and stakeholder identification technologies using the International Standard Content Code (ISCC) and the Creative Passport, while maintaining compatibility with existing and other emerging standards for content identification, we have set a firm foundation on which to build means of discovering, enriching, marketing, and using content with effective application of content metadata through trusted bindings.

By providing trustworthy means of binding such identifiers along with many different types of metadata and rights assertions, the provenance, authenticity, and compliance of content with agreements enable the automated distribution of content and associated rights compensation.

We have focused on both the production of reliable Rights Management Information and just as well on interfaces and tools for the discovery and use of that Information in our Framework, so that new services and businesses can emerge to make Rights Management Information and associated content more useful and available.

Then, we have shown that the DisTRi Framework can be effectively used to help solve some of the biggest emerging problems with digital content. This includes the creation and distribution of inauthentic, misrepresented, mis-contextualized, and unauthorised content.

Finally, we note that to do all of this, we have searched for highly scalable means for each of the basic capabilities we have presented. For example, the use of open content identifiers, Self-Sovereign Identity, and customisable, independently operated ledgers allows many more actors to conceive and provide useful services. We have shown that the DisTRi Framework also scales due to the active use of a two-sided model for information using efficient data ingestion services for recording data and assertions, and derivative services that precompute derivative data for specific applications that use data.

15 About the authors

Albhy Galuten (albhy@galuten.com) is a technology strategist helping creators connect the dots of rights management, working on standards, interoperability, and new development.

Paul Jessop (paul@countyanalytics.com) is a standards and metadata expert seeking to apply technology, industry and copyright knowledge to new situations in the music industry and beyond.

Jack Lacy (lacy@intertrust.com) and David P. Maher (dpm@intertrust.com) are respectively Head of Standards Activities and Chief Scientist at Intertrust Technologies, developing the trusted immutable distributed assertion ledger system.

Titusz Pan (tp@craft.de) is Chief Technology Officer at the ISCC Foundation, promoting open standards helping individuals and organisations better manage digital content.

Philippe Rixhon (prixhon@digiciti.com) is Research and Innovation Director at Digiciti, blueprinting the copyright infrastructure and media marketplaces of the future.

Mark Simpkins (mark@creativepassport.net) is Chief Technology Officer at The Creative Passport, bringing the vision of self-sovereign identity both to life and to market.

16 References

[1] Rixhon, P. (2020). “New media business models to emerge from the IoV”. The Internet of Value. Springer.

[2] Takagi, S. (2020). “Solving challenges in the media sector with DLT”. The Internet of Value. Springer.

[3] European Union (2019). “Directive (EU) 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market”.

[4] 115th Congress of the United States of America (2017-2018). “H.R.1551 - Orrin G. Hatch-Bob Goodlatte Music Modernization Act”.

[5] European Commission (2020). “A European strategy for data”. COM(2020) 66

[6] Rixhon, P. (2020). “The copyright infrastructure, a common European data space”. Conceptual Notes. Academia.

[7] German Federal Ministry for Economic Affairs and Energy (2020). “GAIA-X: Driver of digital innovation in Europe”.

[8] Pólvora A. et alia (ed) (2020). “Scanning the European ecosystem of Distributed Ledger Technologies for social and public good”. European Commission, Joint Research Centre. EUR 30364 EN.

[9] International Federation of Library Associations and Institutions (2009). “Functional Requirements for Bibliographic Records”. https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

[10] Rust, Godfrey and Bide, Mark (2000). “The indecs framework - Principles, model and data dictionary”. https://www.doi.org/factsheets/indecs_factsheet.html

[11] World Intellectual Property Organization. “WIPO-Administered Treaties”. https://www.wipo.int/treaties/en/

[12] European Community (2001). “Directive 2001/29/EC of the European Parliament and of the Council on the harmonisation of certain aspects of copyright and related rights in the information society”. Article 7 §2.

[13] European Union (2014). “Directive 2014/26/EU of the European Parliament and of the Council on collective management of copyright in musical works for online use in the internal market”. Article 24.

[14] 115th Congress of the United States of America (2017-2018). “H.R.1551 - Orrin G. Hatch-Bob Goodlatte Music Modernization Act”.

[15] Treloar, A. “The Five Persistences - which one?”. https://andrew.treloar.net/research/diagrams/five_persistences.pdf.

[16] Statista J. Clement (2020). "Global digital population as of July 2020."

[17] Jinda-Apiraksa, Amornched, Vassilios Vonikakis, and Stefan Winkler. "California-ND: An annotated dataset for near-duplicate detection in personal photo collections." 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX). IEEE, 2013.

[18] Pan, T. “A Proposal for a Modern and Open Content-Based Identifier” (2018), https://iscc.codes.

[19] Hürsch, Walter L., and Cristina Videira Lopes. “Separation of concerns.” (1995).

[20] Charikar, Moses S. “Similarity estimation techniques from rounding algorithms.” Proceedings of the thirty-fourth annual ACM symposium on Theory of computing. 2002.
[21] Ruder, Sebastian, Ivan Vulić, and Anders Søgaard. “A survey of cross-lingual word embedding models.” Journal of Artificial Intelligence Research 65 (2019): 569-631.
[22] Zauner, Christoph. “Implementation and benchmarking of perceptual image hash functions.” (2010).
[23] Bhatia, Rhythm, et al. “Analysis of audio features for music representation.” 2018 7th international conference on reliability, infocom technologies and optimization (trends and future directions) (ICRITO). IEEE, 2018.
[24] Paschalakis, Stavros, et al. “The MPEG-7 video signature tools for content identification.” IEEE transactions on circuits and systems for video technology 22.7 (2012): 1050-1063.
[25] Xia, Wen, et al. “Fastcdc: a fast and efficient content-defined chunking approach for data deduplication.” 2016 {USENIX} Annual Technical Conference ({USENIX}{ATC} 16). 2016.
[26] Saxena, Priyamvada. “Analysis of Various Hash Function.”
[27] Broder, Andrei Z., et al. “Min-wise independent permutations.” Journal of Computer and System Sciences 60.3 (2000): 630-659.
[28] Merkle, Ralph C. “A digital signature based on a conventional encryption function.” Conference on the theory and application of cryptographic techniques. Springer, Berlin, Heidelberg, 1987.
[29] Chevet, Sylve. “Blockchain technology and non-fungible tokens: Reshaping value chains in creative industries.” Available at SSRN 3212662 (2018).
[30] “Working for development, integration and adoption of Self-Sovereign Identities (SSI) technologies”. https://essif-lab.eu/.
[31] IETF OAuth Working Group. “OAuth 2.0”. https://oauth.net/2/.
[32] Williams, C. “GlobalSign screw-up cancels top websites' HTTPS certificates”. https://www.theregister.com/2016/10/13/globalsigned_off/. (2016).
[33] “X.509 Security”. https://en.wikipedia.org/wiki/X.509#Securit.
[34] Grimes, Roger A. “4 fatal problems with PKI”. https://www.csoonline.com/article/2942072/4-fatal-problems-with-pki.html. (2015).
[35] Santesson, S., Myers, M., Ankney, R., Malpani, A., Galperin, S., Adams, C. “X.509 Internet Public Key Infrastructure Online Certificate Status Protocol - OCSP”. https://tools.ietf.org/html/rfc6960. (2013).
[36] Pettersen, Y. “The Transport Layer Security (TLS) Multiple Certificate Status Request Extension”. https://tools.ietf.org/html/rfc6961**.** (2013).
[37] Egberts, Alexander. "The Oracle Problem-An Analysis of how Blockchain Oracles Undermine the Advantages of Decentralized Ledger Systems." Available at SSRN 3382343. (2017).
[38] Mejía, P. “The Success Of Streaming Has Been Great For Some, But Is There A Better Way?” https://www.npr.org/2019/07/22/743775196/the-success-of-streaming-has-been-great-for-some-but-is-there-a-better-way. (2019).

[39] Ginsberg, Jane C. “Losing Credit: Legal Responses to Social Media Platforms’ Stripping of Copyright Management Metadata from Photographs”. https://www.mediainstitute.org/2016/05/30/losing-credit-legal-reponses-to-social-media-platforms-stripping-of-copyright-management-metadata-from-photographs/. (2016).

[40] Deahl, D. “Metadata Is the Biggest Little Problem Plaguing The Music Industry”. https://www.theverge.com/2019/5/29/18531476/music-industry-song-royalties-metadata-credit-problems. (2019).

[41] Smith, E. “Songwriters Lose Out on Royalties”. https://www.wsj.com/articles/songwriters-lose-out-on-royalties-1444864895. (2015).

[42] Wactlar, H. D., Christel, M. G. “Digital Video Archives: Managing Through Metadata”. https://www.clir.org/pubs/reports/pub106/video/.

[43] Wang, Shuai, et al. "Blockchain-enabled smart contracts: architecture, applications, and future trends." IEEE Transactions on Systems, Man, and Cybernetics: Systems49.11 (2019): 2266-2277.

[44] Maher, David. "On Software Standards and Solutions for a Trusted Internet of Things." Proceedings of the 51st Hawaii International Conference on System Sciences. 2018. Available at: https://scholarspace.manoa.hawaii.edu/handle/10125/50599.

December 2020