Trustworthy preservation and access metadata using distributed ledger technology (DLT)

Sven Schlarb, Roman Karl

AIT Austrian Institute of Technology, Vienna, Austria

ABSTRACT

Trusted digital repositories require maintaining the integrity and authenticity of digital objects throughout their lifecycle. Digital signatures can establish trust regarding events such as submission, dissemination, and archiving. The PREMIS metadata standard is commonly used for recording preservation information, and agents performing events related to digital objects. Our approach utilizes distributed ledger technology (DLT) to record PREMIS entities in a trustworthy manner, following European eArchiving Initiative standards. This enhances trust in electronic archiving by providing auditable preservation and access metadata. PREMIS metadata can also define rights and certificates for OAIS functions, and integration with identity services enables access authorization. A software ontology or taxonomy is needed for compliance with GDPR, and DLT can provide GDPR-compliant access to repository data, crucial for scientific purposes such as data science and big data.

Keywords: Trusted Archiving, Digital Preservation, Blockchain, DLT

I. INTRODUCTION

Trusted digital repositories need to maintain the integrity and authenticity of digital objects throughout their information lifecycle, including submission, archiving, maintenance, and delivery processes. Digital signatures can be used to establish trust in these events that change the state of a repository. For instance, during submission, an agent can sign an information package to identify themselves as the author of the submission. In the context of dissemination, digital signatures can serve to identify the source of the information package. Furthermore, for archiving of digital objects, digital signatures can be employed to ensure their integrity and authenticity over the long term.

Considering the long-term preservation measures that may need to be applied to digital objects, information packages are subject to a lifecycle. Policy and restrictions related to events that change the state of information packages must be respected, and the author and rationale of such measures must be documented. PREMIS, a widely used metadata standard in the archival and library domain, facilitates the recording of preservation information. Agents, who can be individuals or software, trigger events related to digital objects, such as intellectual entities, representations, or bit streams, and there is the option to define rights pertaining to these digital objects.

Against this background, we present an approach that leverages distributed ledger technology (DLT) to record evidence of the existence of central PREMIS entities in a

trustworthy manner. In accordance with the standards recommended by the European eArchiving Initiative[1], we describe a specific implementation that enables the transparent, traceable, and tamper-proof recording of the information package lifecycle. The aim of this approach is to enhance trust in electronic archiving and digital repositories by providing an interoperable and cost-efficient means to create system-independent, auditable preservation metadata.

In addition to documenting the information package lifecycle, PREMIS preservation metadata can also be used to define rights regarding use or manipulation of digital objects for the central OAIS (Open Archival Information System) functions of submission, archiving, and dissemination. Through integration with central identity services, access authorization can be verified based on rights and certificate data. This enables the modeling of permissions for ordering and accessing repository objects.

Furthermore, we highlight the need for a software ontology or taxonomy in the field of long-term preservation, with the goal of ensuring compliance with guidelines such as the General Data Protection Regulation (GDPR) pertaining to the processing of data within a repository. The decentralized provision of authorization, proof of authenticity, and certificates using distributed ledger technology (DLT) forms the foundation for GDPR-compliant access to repository data, which is crucial for scientific purposes, especially in the areas of data science and big data.

II. RELATED WORK

Several standards exist for ensuring trusted archiving. The Reference Model for an Open Archival Information System (OAIS) outlines the requirements for archives or repositories to maintain long-term preservation of digital information. Building on this reference model, various initiatives have produced recommendations for certification criteria related to trustworthy repositories, as cited in [4], [5], and [3]. While these publications primarily focus on organizational infrastructure aspects and do not delve into the technical means for building trust, they provide a general framework for constructing 'accountable record-keeping systems' [4, p. 8].

The significance of blockchain technology in archiving is evident from the numerous publications on this topic. Our approach closely aligns with the model of a blockchain-based system for facilitating the long-term preservation of digitally signed records, as presented in [6] and [7]. However, our approach distinguishes itself by presenting a generic use case that is applicable to any type of archive and proposes a way to link preservation and access rights metadata with the blockchain registry.

III. AUDITABLE PROVENANCE AND ACCESS

We present basic metadata elements to create a system-independent and auditable preservation and access rights record. This may concern rights to perform preservation actions, rights to process metadata to make objects discoverable in a catalogue system or the rights to perform operations, such as Optical Character Recognition (OCR) or Full-Text-Indexing with the purpose to enable search and retrieval of the digital objects, for example. It may also concern the rights of persons who have requested access to information objects which requires accepting specific conditions

When performing any of these operations, the operator must be identifiable by the system. If the operation involves more than one agent, any party involved needs to be identified.

In the following we consider the following basic requirements regarding the use of rights statements.

- It must be possible to define rights statements on information package, representation, or file level given that identifiers for the corresponding components are available.
- It must be possible to define rights statements related to the application of software agents enabling information retrieval, such as Optical Character Recognition (OCR) or full-text indexing software, for example.
- It must be possible to define rights statements granted to specific information applicants.

PREMIS defines three kinds of objects:

- File – a computer file, such as like a PDF or JPEG file.
- Representation - set of all file objects needed to render an Intellectual Entity.

[1] https://digital-strategy.ec.europa.eu/en/activities/earchiving

- Bitstream – a set of data units (bits) within a file that has meaningful common properties for preservation purposes, such as the audio bitstream in an AVI container, for example

Additionally, since PREMIS version 3.0 [2] there is the option to model Intellectual Entities as PREMIS objects. Access rights can therefore be expressed using in relation to the information package, representation, or file level. For the information package the intellectual entity object is used which in the scope of the container relates to a concrete version of the information package. During the life cycle of the archival information package representations may be added if preservation actions are taken, such as the conversion of a file format, for example. As illustrated in Figure 1, PREMIS elements are used to record elements for the intellectual entity and representations. Using the PREMIS `relationshipType = derivation` and the `relationshipSubType = derived from` it is possible to document the provenance of representations during the life-cycle and build a dependency graph of preservation operations.

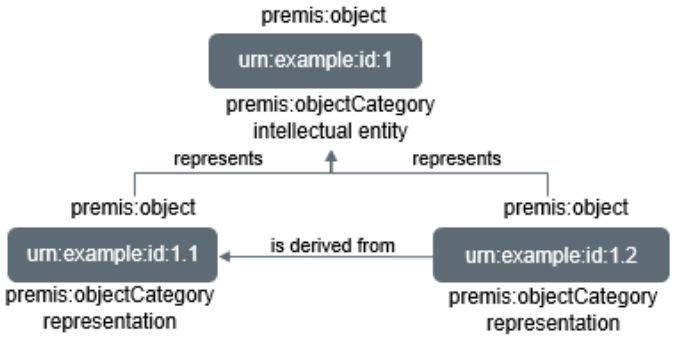

Figure 1 PREMIS elements

Operators changing the status of or requesting access to archived information are represented in form of PREMIS agents. This way it is possible to allow or prohibit operations or access to objects at any of the defined organisational levels (intellectual entity, representation, bit stream).

By using a software taxonomy, such as the illustrative example shown in Figure 2, permissions can be assigned to a category of software tools at any level of a software taxonomy.

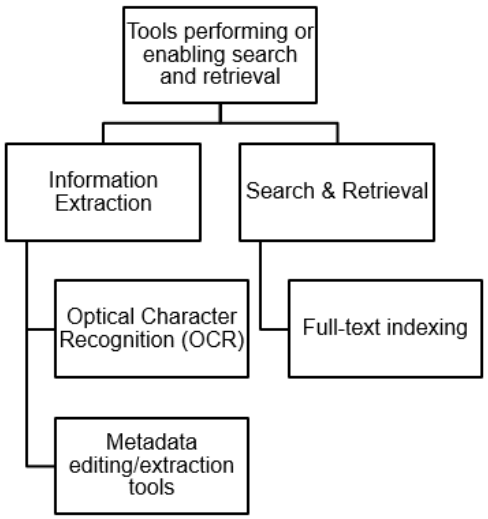

Figure 2 Illustrative example of a software taxonomy

When it comes to requests for access to archive objects, the contractual agreement between agents can be documented in PREMIS - anonymously if necessary. In this case, the blockchain transactions serve to prove the existence of documents signed by the contracting parties.

To ensure an autitable record of preservation actions, evidence of agents and events are recorded in the blockchain. This way the record of preservation and access metadata is done a system-independent and auditable way.

IV. DLT PROTOTYPE IMPLEMENTATION

To store data in the blockchain, or more precisely, to execute a transaction that will be documented in a block, an account within the blockchain system is necessary. This account may represent either an individual user or an external system that manages multiple user accounts. In addition to a digital account secured by public-key cryptography, a connection to a specific person or organization must be established, particularly when legal contracts are involved. Our emphasis will be on individuals, commonly referred to as "natural persons", as many relevant legal agreements are made at this level. The proof-of-concept [3] implements a user management that does not include an official verification of an identity by the respective national agencies of EU states. But the architecture is modular enough to allow such an extension without large code

[2] http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

[3] https://github.com/E-ARK-Software/blockchain-notary-poc

changes. A potential extension could integrate eID[4], a digital building block that originated also from the CEF program, which takes care of cross-border verification of identities. Similarly, the European Self-Sovereign Identity Framework (ESSIF) built into EBSI can facilitate the generation of user accounts with verified identities based on official documents and make it usable with the rest of the EBSI functionality. For the proof-of-concept we won't be able to reach this level of security on this task, but on the other hand have a higher level of flexibility on creating the accounts. Hereby, the security lies in the hand of the system operator which might not reach the same level of trust as a national agency.

Giving access to data is also not a task that is typically solved only by a smart contract. First, because most blockchain systems have no sophisticated reading protection, which means that data is readable by default for parties with access to the system. But a blockchain is often not an ideal storage system for many kinds of data, because of its persistent nature and its reduced capacity and throughput. Therefore, archival content will clearly not be stored on the blockchain but the blockchain holds the proof of existence of access permissions, for example. Thus, enforcing the rules of the blockchain will be done by a component outside the blockchain system. This component does not need a distributed architecture and can be located at the archive. For multiple archives the component can simply be run as multiple instances, or an archive could use its own implementation if it wishes so. It is important to note that by doing so, we do not lose the advantages of the blockchain system. If an archive gave access sensitive archived information without having a legal contract established on the blockchain, this would be in its own control and responsibility. On the other hand, an archive that does not give access to an applicant even though a legal contract has been established would violate that agreement, which could by proven by the applicant.

The European Blockchain Services Infrastructure (EBSI)[5] is a cooperation of 29 countries and the European Commission. It is a private blockchain-based system that is rather small in the number of nodes with about 30 nodes. On the other hand, it is quite large in the sense that it spans the continent with several countries and different legislations, is built on a complex stack, and provides different services. It largely builds upon the Ethereum ecosystem with several smart contracts written in Solidity, that define the core of the EBSI functionality. This private Ethereum network is not accessible directly from outside by users and external developers. Instead, there are several higher-level APIs as the only way to access the system from outside. Developers can use this API to write decentralized applications in a similar way as with interacting with custom smart contracts directly but with the additional EBSI compatibility. As EBSI is an already established infrastructure and furthermore provides useful services like a Europe-wide standardized identification, it provides an additional value for auditable provenance.

One of EBSI's APIs is particularly interesting for referencing to data entries to prove their existence and validity at a certain time. It is called the Timestamp API and basically stores hashes and associates them to the timestamps at the point of creation of the entry. The hashing algorithm is not fixed by the Timestamp API but can be picked by the user out of a list of standardized algorithms.

The timestamp is added by the EBSI system when the entry is written to the blockchain. The original data can be stored off-chain. That might be possible also via the EBSI infrastructure or outside of it via a separate application. Unfortunately, at the time of writing, due to its early stage the ways of accessing the EBSI are still very limited and restricted, but this might change in the future.

An important aspect of the dissemination is the definition and the agreement of the usage rights.

When a DIP containing an OCR PDF representation was created, it can be made available for access with certain restrictions. This is a crucial part, because historical documents can contain sensible information, in particular personal data. To define in which way and to what extent a DIP can be used, a legally binding contract between two parties, the provider, and a requester, must be put in place. In the context of systems that use a blockchain as basic data structure, there are pieces of code, called smart contracts, which are sometimes seen as an automated form of a legal contract. But this is only

[4] https://ec.europa.eu/digital-building-blocks/wikis/display/DIGITAL/eID

[5] https://ec.europa.eu/digital-building-blocks/wikis/display/EBSI

true in certain cases because smart contracts can only control very specific aspects of a legal contract. A smart contract cannot prevent the requester to use the DIP in any way that would violate the legal contract. But still, having a blockchain as data structure where it is not possible to modify existing entries helps us to digitally record an agreement between two parties and put the legally binding contract in place. The central part of the legal contract is the text that describes the legal aspects, which we call “license”.

The process of recording the agreement on the blockchain consists of five steps which are shown in

Register dissemination representation for access. The DIP is identified and referenced to by a UUID. The DIP itself is never put on the blockchain, but only its identifier.

Create text license document. Most of the time, we will deal with a small set of standard licenses, but it is also possible to assemble a license out of a set of standard clauses or to set up an individual license. A license is hashed to prohibit future modifications and the hash is used as identifier for a license on the blockchain. Like the DIP, the license itself won’t be stored on the blockchain.

Provider assigns license to dissemination representation. An entry on the blockchain is made to record the bundling of the DIP with a specific license. Everyone with access to the system will then be able to see the available offers. Since the blockchain contains only the identifiers, a customer will need not only the information from the blockchain to assess the offer. What is important, is that the data on the blockchain is sufficient for a customer to verify the validity of the offer.

Requester accepts license. The requester is the first to sign the offer via an entry on the blockchain.

Provider approves request. The signature of the provider via an entry on the blockchain finalizes the legally binding contract between one requester and provider for one DIP.

V. CONCLUSION

The concepts and approach presented in this article can be implemented with Distributed Ledger Technology or Blockchain services which offer a function for registering a hash value and providing a timestamp as return value. These minimum requirements will allow tracing the creation and integrity of information objects. To also provide evidence for the authenticity of information objects, the events need to be linked to the account. If the requirement is to know about real identities behind the accounts, further identification services, such as the European eID services, need to be integrated.

Additionally, the implementation of the security concept allows enforcing the access conditions defined by a license which is assigned to the information objects and which, in case of restricted information resources, must be accepted by the information requestor before using the information package is granted.

VI. ACKNOWLEDGEMENT

This project has received funding from the European Union’s CEF programme with action No 2020-EU-IA-0185 under grant agreement No INEA/CEF/ICT/A2020/2397190.

VII. REFERENCES

[1] K. Azbeg, O. Ouchetto, and S. J. Andaloussi. Access control and privacy-preserving blockchain-based system for diseases management. IEEE Transactions on Computational Social Systems, pages 1–13, 2022.

[2] D. Di Francesco Maesa and P. Mori. Blockchain 3.0 applications survey. Journal of Parallel and Distributed Computing, 138:99–114, 2020.

[3] D. R. C. T. Force. Trustworthy repositories audit certification: Criteria and checklist. Research libraries group (rlg), mountain view, ca, 2007, Technical report, RLG-NARA, 2007.

[4] R.-O. W. G. on Digital Archive Attributes. Trusted digital repositories: Attributes and responsibilities. Technical report, OCLC, 2002.

[5] W. G. on Digital Archive Attributes. An audit checklist for the certification of trusted digital repositories. RLG-OCLC, 2005.

[6] Bralić, V., Stančić, H. and Stengård, M. (2020), "A blockchain approach to digital archiving: digital signature certification chain preservation", Records Management Journal, Vol. 30 No. 3, pp. 345-362. https://doi.org/10.1108/RMJ-08-2019-0043

[7] Stančić H, Bralić V. Digital Archives Relying on Blockchain: Overcoming the Limitations of Data Immutability. Computers. 2021; 10(8):91. https://doi.org/10.3390/computers10080091