RESEARCH DATA MANAGEMENT AND LONG-TERM PRESERVATION USING BLOCKCHAIN

Dr. **Hrvoje Stančić**, full professor
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb, Croatia
hstancic@ffzg.hr

Zagreb/Geneva, 22 October 2020

Contents

1. Introduction

- Documents, records and research data today – increasingly
  - created, analysed, used, reused

  in the digital form

- Requirements for the (long-term) preservation (LTP) of digital resources in light of constant change and development of ICT
  - LTP actions = conversion, migration, emulation, virtualization

1. Introduction …

- LTP challenges – how to preserve
  - authenticity
  - integrity
  - reliability
  - usability
  - non-repudiation
  - security
  - confidentiality
  - proof of ownership

⇒ Trustworthiness
  - authenticity, accuracy, reliability

- Concepts arising from archival theory

1. Introduction … TRUSTWORTHY DATA

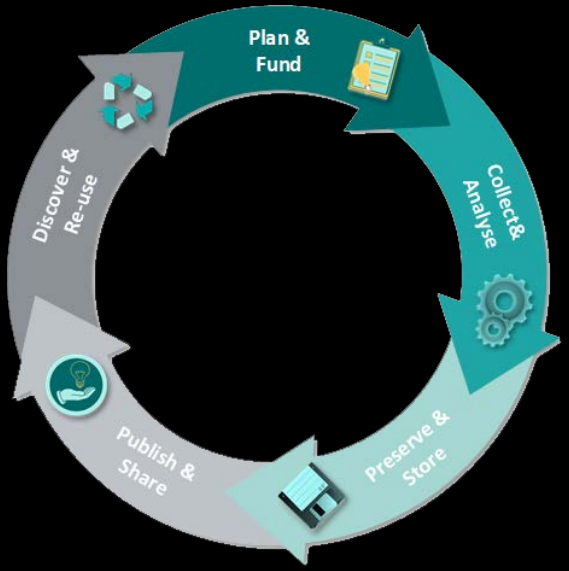
- Authenticity
  - "the quality of a record that is what it purports to be and that is free from tampering or corruption"
- Accuracy
  - "the degree to which data, information, documents or records are precise, correct, truthful, free of error or distortion, or pertinent to the matter"
- Reliability
  - "exists when a record can stand for the fact it is about and is established by examining the completeness of the record's form and the amount of control exercised on the process of its creation"

(Multilingual Archival Terminology, http://www.ciscra.org/mat)

1. Introduction …

- Research data lifecycle
  - creation
  - archiving
  - publication
  - preservation
  - reliable (re)use & attribution

- Is research data safely preserved?



1. Introduction …

- The 'Bit List' of Digitally Endangered Species – 2019 report, Digital Preservation Coalition (DPC), https://www.dpconline.org/digipres/champion-digital-preservation/bit-list

–	Research data published through repositories	Published research data appended to journal articles Semi-Published Research Data	Unpublished research data from US Government researchers (archival responsibility well developed)	Unpublished research data (intentionally)

1. Introduction …

- Intellectual property rights (IPR)
  - licences – only the rights holder (or someone with a right or licence to act on their behalf) can grant a licence
  - establish IPR pertaining to the data before any licensing takes place
  - nature of a licence is to *expand* rather than *restrict* what a licensee can do, some licences are presented within *contracts*, and contracts *can* place additional restrictions
  - waivers – giving up one's rights to a resource → infringement becomes a non-issue, but only the entity holding the rights can waive them

Ball, Alex (2014). 'How to License Research Data'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: https://www.dcc.ac.uk/guidance/how-guides/license-research-data

1. Introduction …

- Challenges
  - proving data ownership
  - verification of research results
  - establishing data provenance
  - automation of licencing contracts

- Can blockchain help?

2. InterPARES Trust project

- Trust and Digital Records in an Increasingly Networked Society (2013-2019)
  - led by Luciana Duranti
  - https://interparestrust.org
  - 499 researchers
  - 7 teams: North America, Europe, Latin America, Asia, Australasia, Africa, Transnational Team

2. InterPARES Trust – EU study no. 31

Model for Preservation of Trustworthiness of the Digitally Signed, Timestamped and/or Sealed Digital Records (TRUSTER Preservation Model)

- the Team: Hrvoje Stančić (lead), Victoria Lemieux, Natasha Khramtsovsky, Enigio Time AB, Croatian Financial Agency FINA, FHSS GRAs
- a model for blockchain-based digital signatures' Validity Information Preservation (VIP) solution

TRUSTCHAIN

3. Blockchain enabling concepts

1. Hash algorithm
2. Merkle tree
3. Chaining of top hashes
4. Distributed consensus

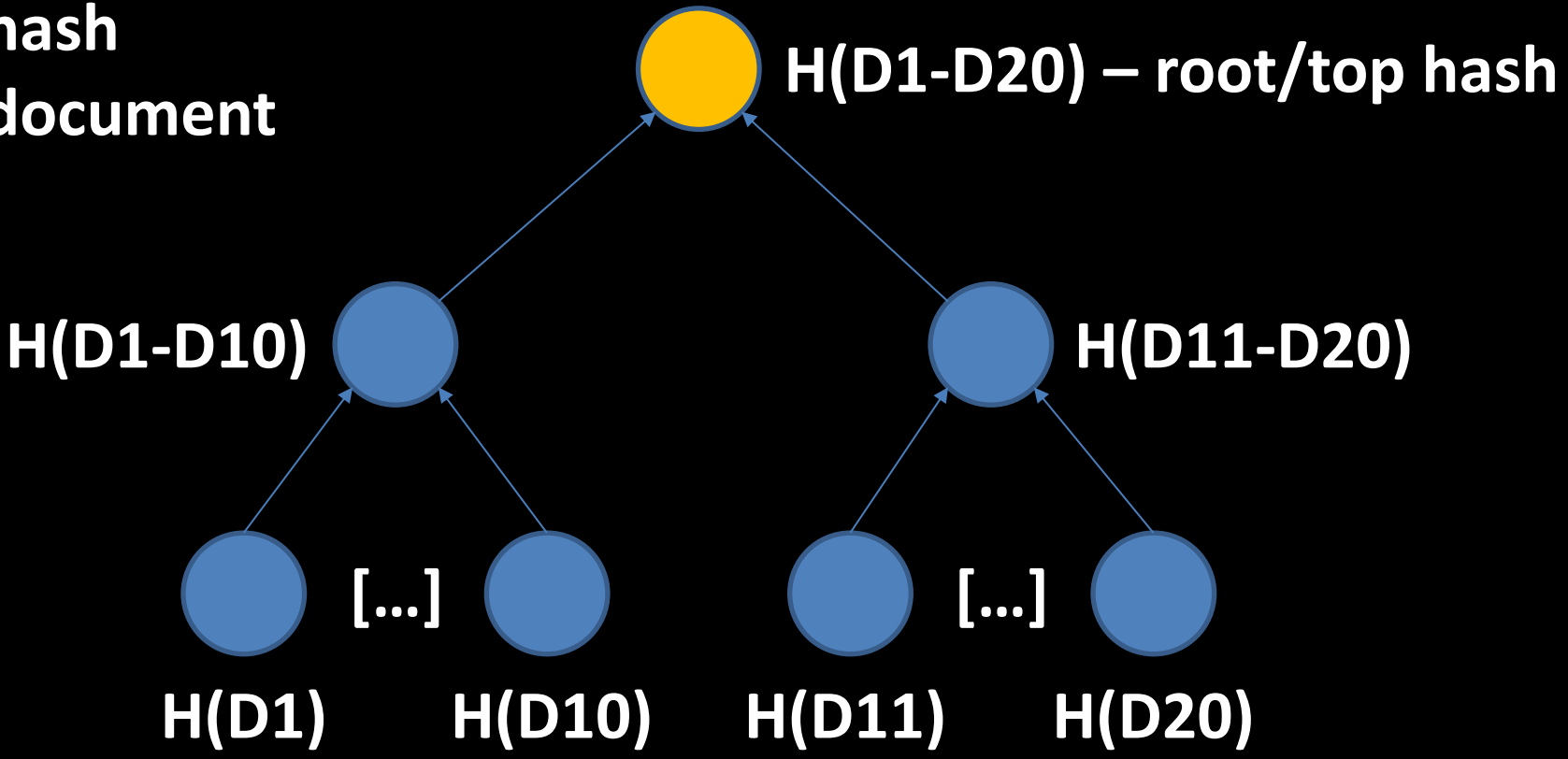
1. Hash algorithm

SHA-256 – example of a hash value of a document

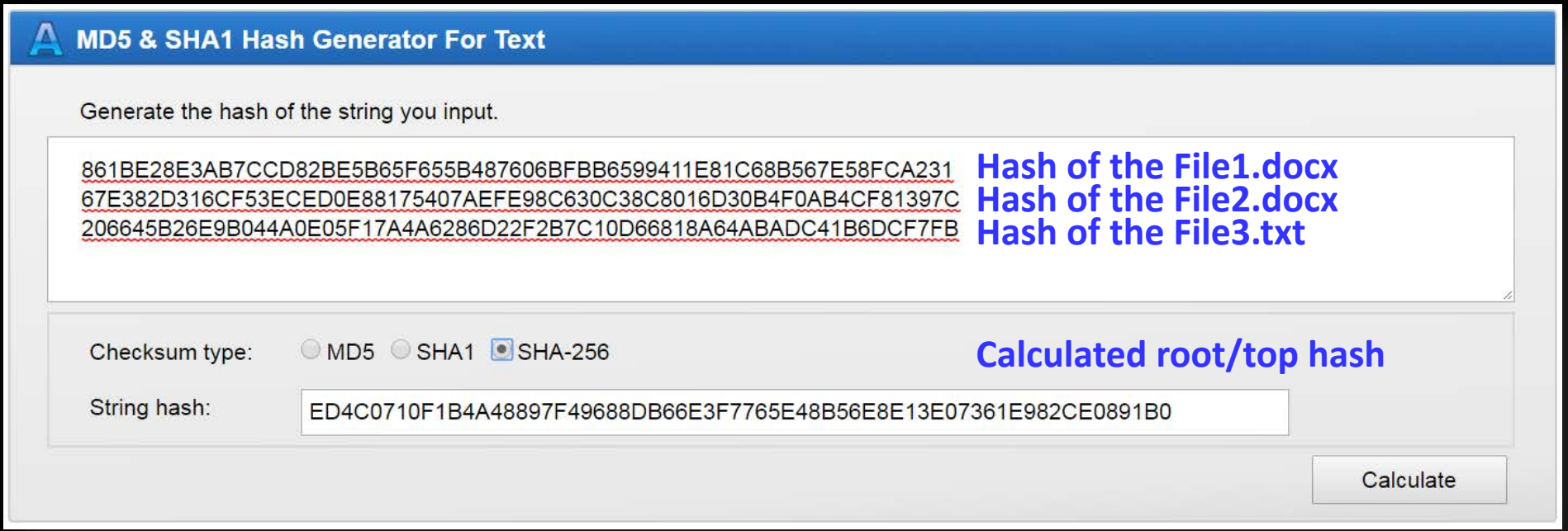
```
7d8c5b62dcb44023
3f7eaac1ec49e4c3
86b8089c37d69ab5
1bc674b8877cb032
```

2. Merkle tree

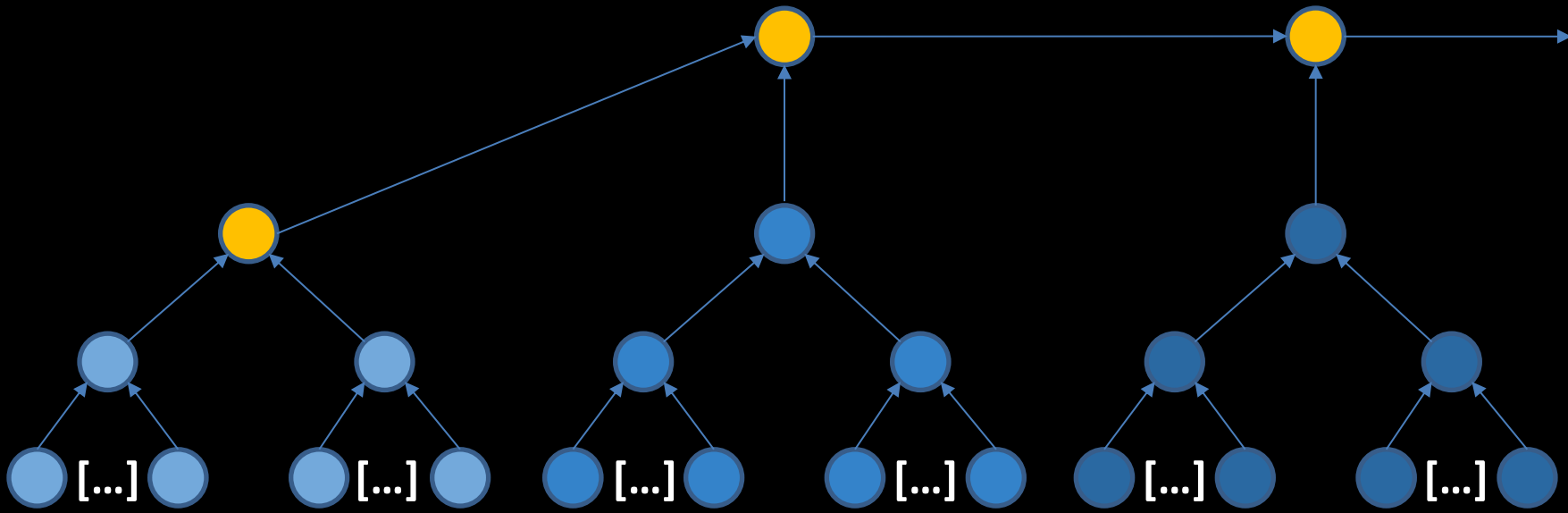
H – hash
D – document



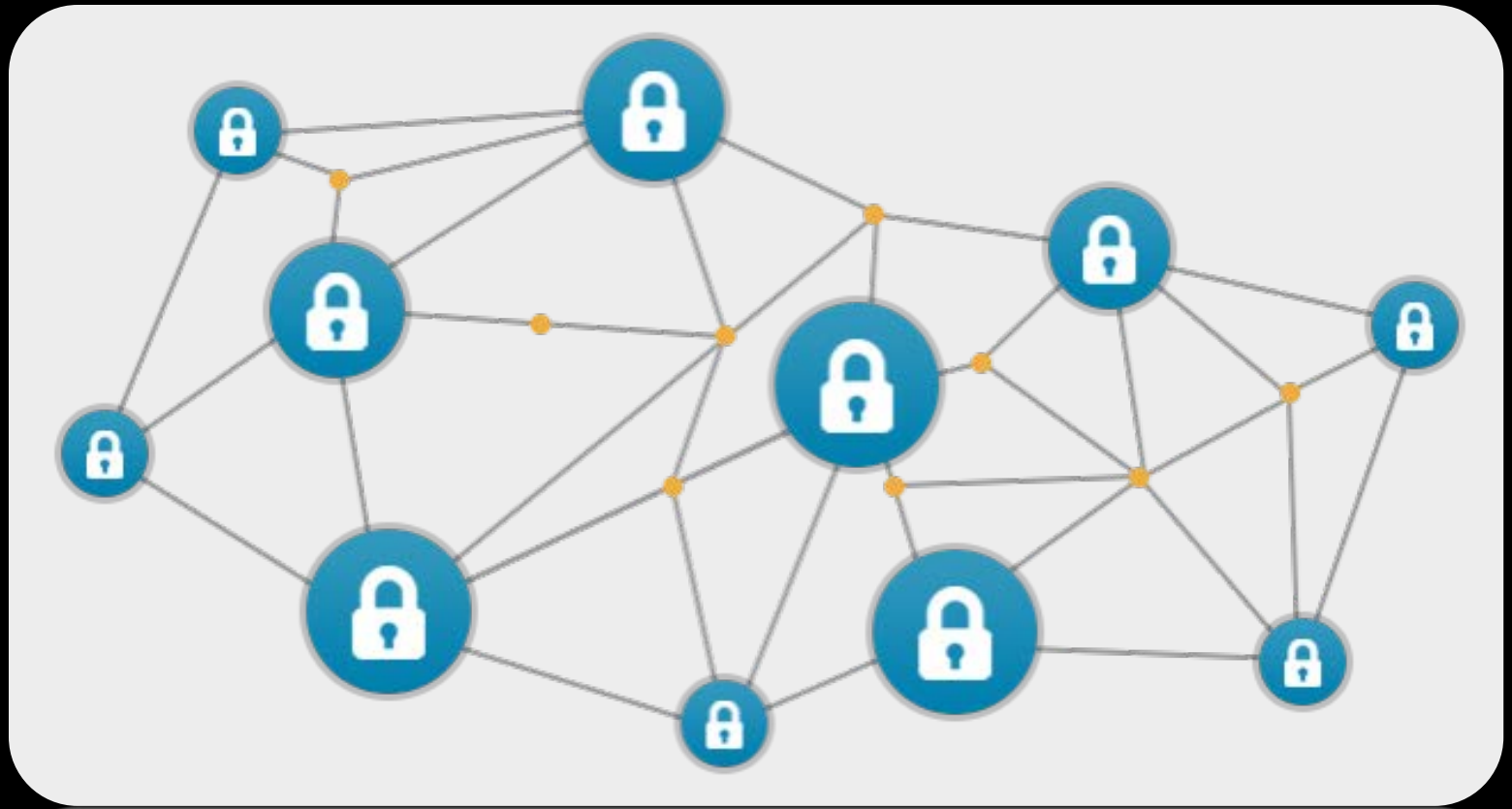
2. Merkle tree



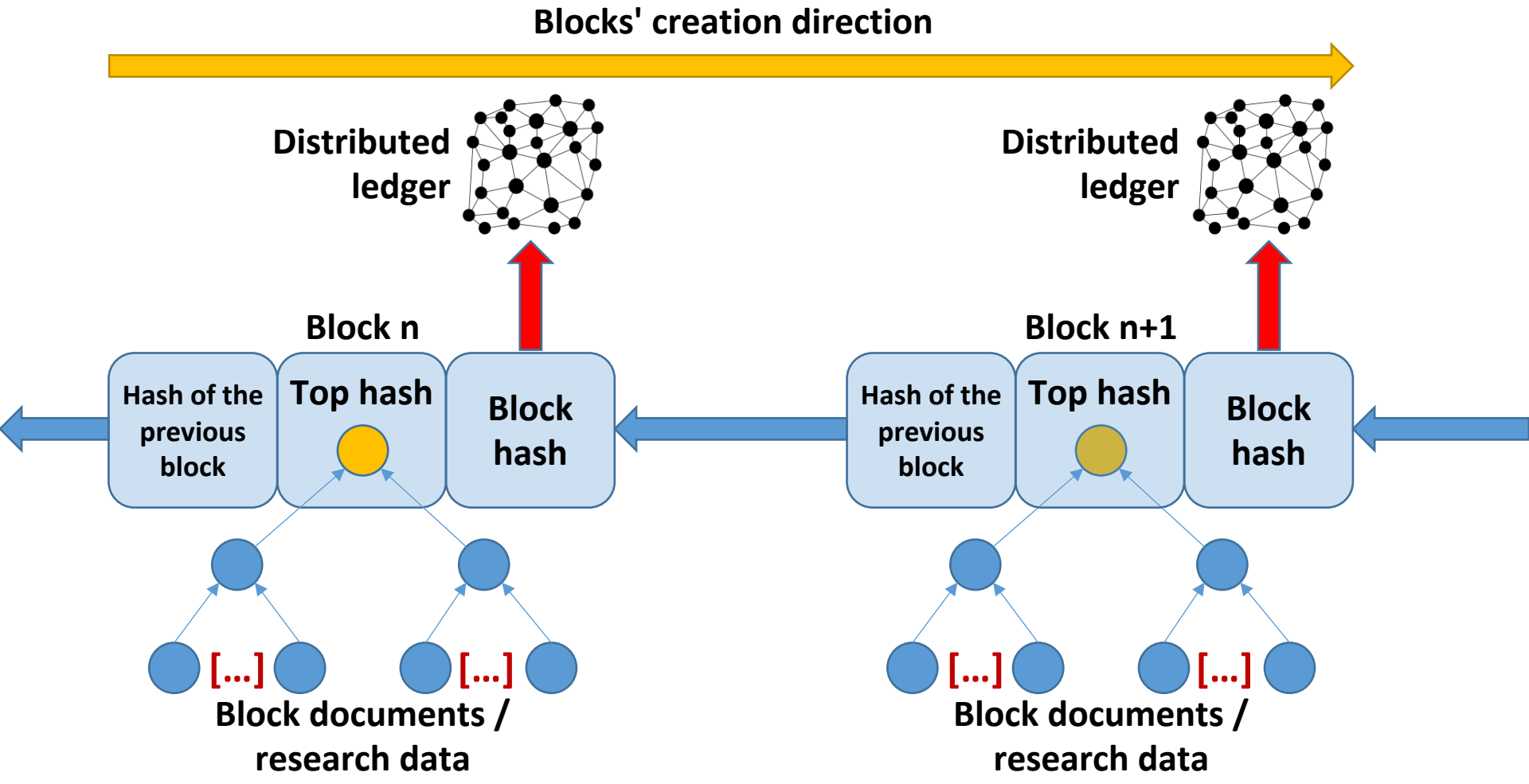
3. Chaining of top hashes



4. Distributed (peer-to-peer) consensus



Blockchain

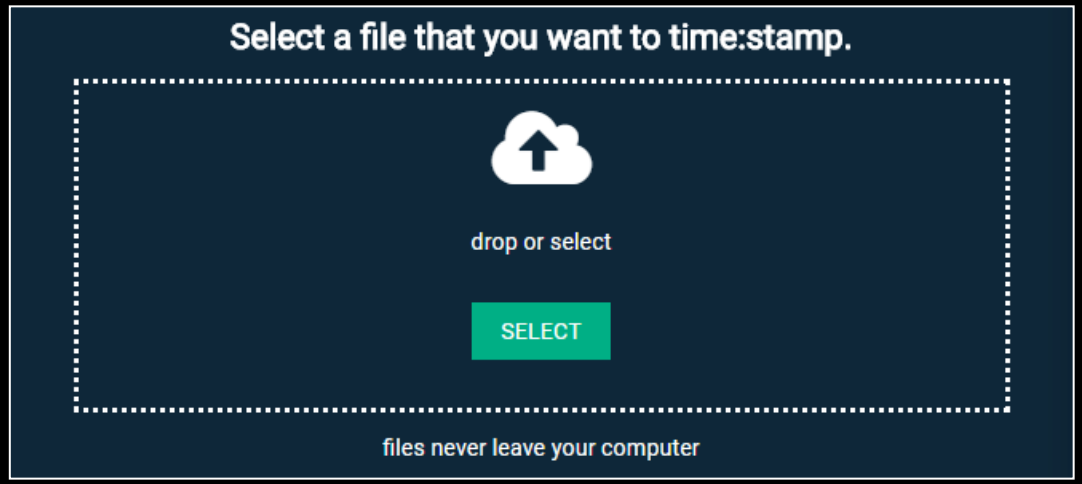

Blocks' creation direction
Distributed ledger
Distributed ledger
Block n
Block n+1
Hash of the previous block
Top hash
Block hash
Hash of the previous block
Top hash
Block hash
[...]
[...]
[...]
[...]
Block documents / research data
Block documents / research data

4. Blockchain and RDM

- Can you prove that a particular research dataset existed at certain point in time (proof of contents)?
- Can it be trusted? Copyright!
- Do you need to significantly change / improve your RDM processes?

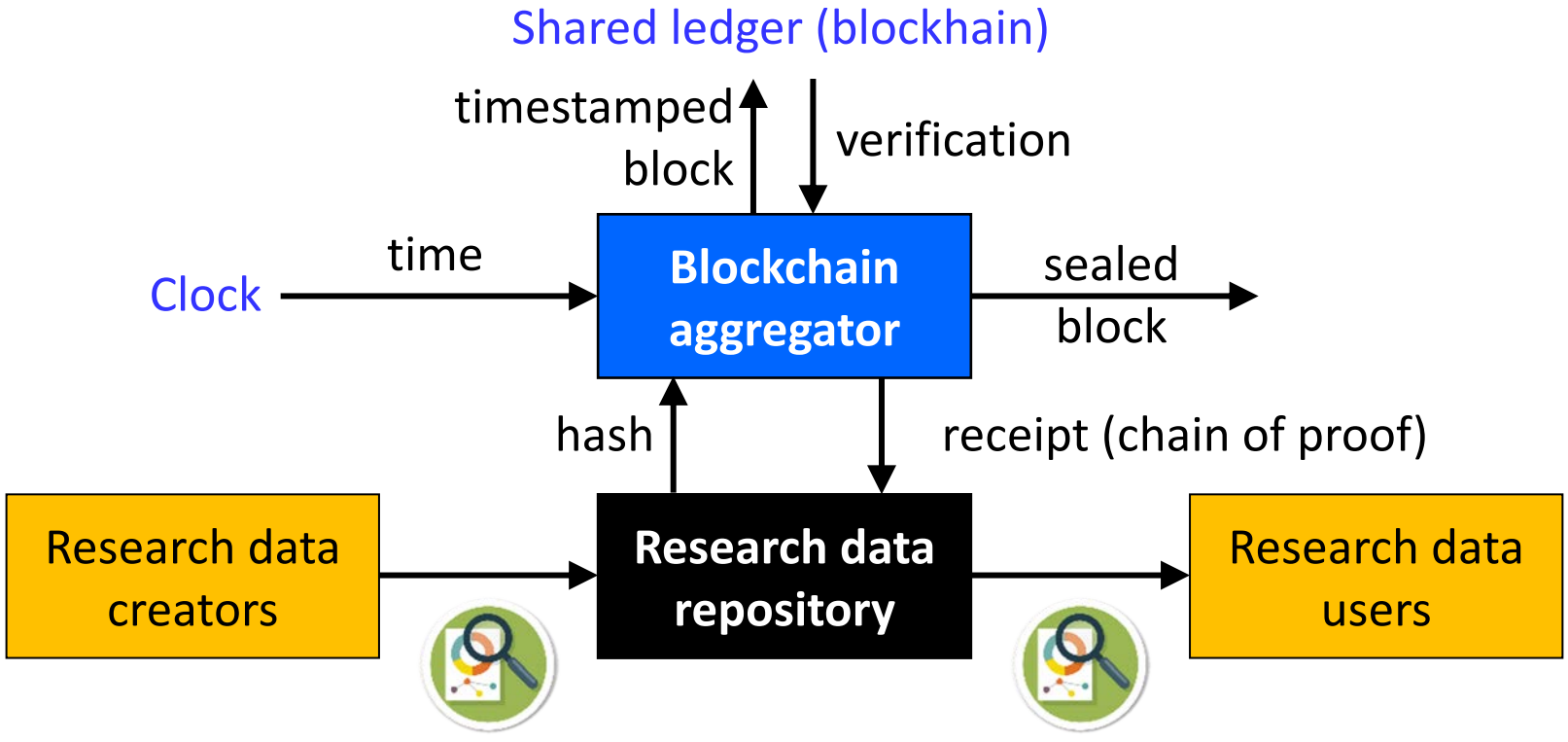
4. Blockchain and RDM …

- EnigioTime – time:beat solution
  - https://timebeat.com/ time:beat by Enigio
  - digital fingerprint
  - reliable timestamp
  - independent verification



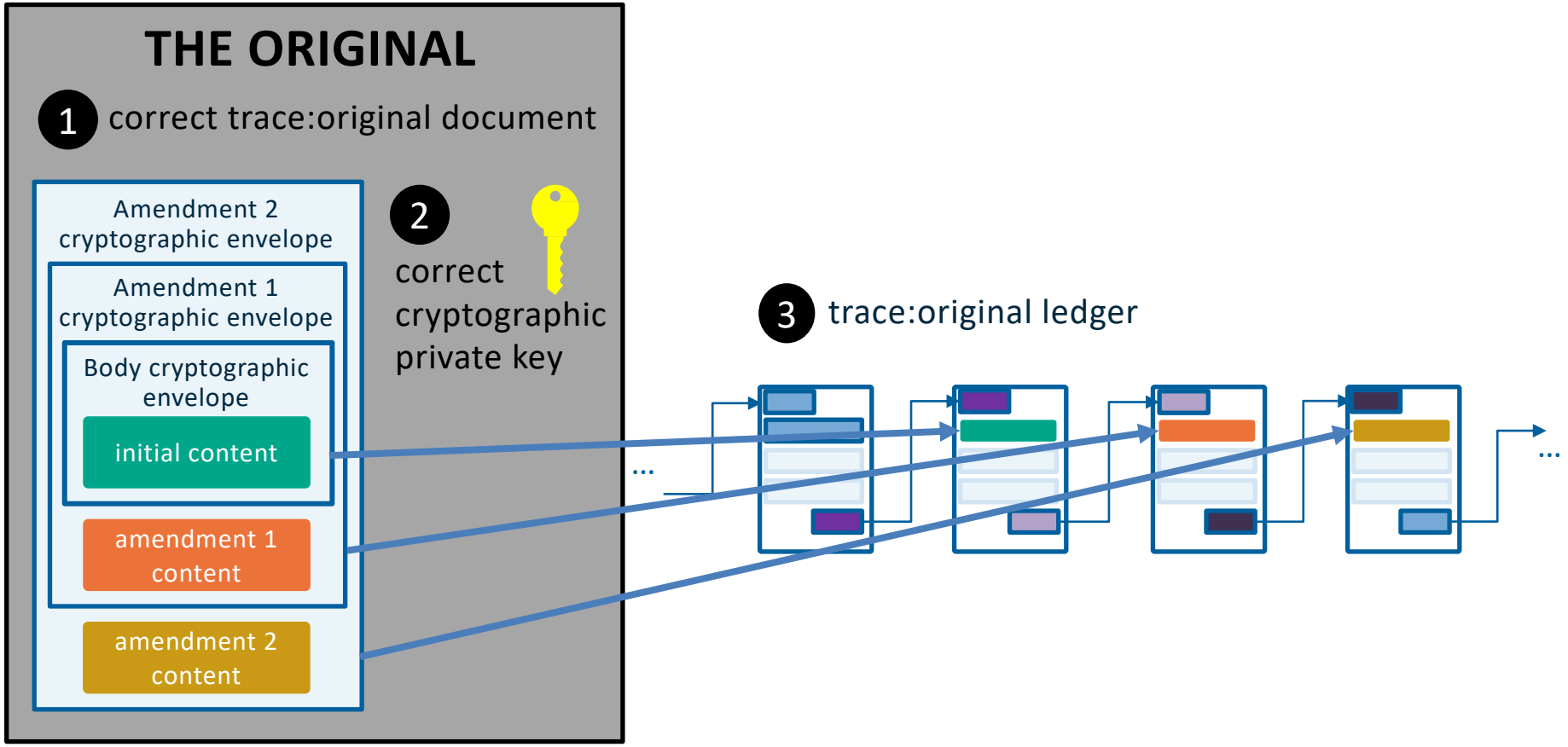
4. Blockchain and RDM …

- time:beat – adding blockchain functionality



4. Blockchain and RDM …

- trace:original – creating a "digital original"



BY USING BLOCKCHAIN …

You can
confirm integrity
of research data.

BY USING BLOCKCHAIN …

You **can add to**
but you **cannot change**
research datasets.

BY USING BLOCKCHAIN …

You can
confirm sequence
of research results.

BY USING BLOCKCHAIN …

You can
keep complete
immutable audit trail.

BY USING BLOCKCHAIN ...

You can

distribute copies

of "digital original"

datasets.

BY USING BLOCKCHAIN …

Anyone can **verify**
if a copy corresponds to the (current) "digital original" datasets.

BY USING BLOCKCHAIN …

You can
prove ownership
and effectively
manage licensing.

7. Conclusion

- Trusted and preserved research data
  - **establish** a new generation of RDM processes in the context of long-term preservation
  - **enable** blockchain functionality through easy-to-connect-to API
  - **preserve** authentic, accurate, and reliable (i.e. trustworthy) digital records / research data with the help of blockchain principles

Resources

- Ball, Alex (2014). 'How to License Research Data'. DCC How-to Guides. Edinburgh: Digital Curation Centre. https://www.dcc.ac.uk/guidance/how-guides/license-research-data
- Bralić, V., Kuleš, M., & Stančić, H. (2017). 'A model for long-term preservation of digital signature validity: TrustChain'. In: I. Atanassova, W. Zaghouani, B. Kragić, K. Aas, H. Stančić, & S. Seljan (Eds.), *INFuture2017: Integrating ICT in Society*, pp. 89-113, https://www.researchgate.net/publication/321171227_A_Model_for_Long-term_Preservation_of_Digital_Signature_Validity_TrustChain
- Enigio Time, https://www.enigio.com/
- InterPARES Trust research dissemination https://interparestrust.org/trust/research_dissemination
  - look for TRUSTER Preservation Model (EU31) documents
- Multilingual Archival Terminology, http://www.ciscra.org/mat
- The 'Bit List' of Digitally Endangered Species – 2019 report, Digital Preservation Coalition (DPC), https://www.dpconline.org/digipres/champion-digital-preservation/bit-list

THANK YOU!

Research data management and long-term preservation using blockchain

Dr. **Hrvoje Stančić**, full professor
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb, Croatia
hstancic@ffzg.hr

LinkedIn

Zagreb/Geneva, 22 October 2020