SMARDY: Zero-Trust FAIR Marketplace for Research Data

Ion-Dorinel Filip[‡], Cosmin Ionite[‡], Alba González–Cebrián[§], Mihaela Balanescu[*], Ciprian Dobre[‡],
Adriana E. Chis[§], Dave Feenan[†], Adrian-Alexandru Buga[‡], Ioan-Mihai Constantin[‡],
George Suciu[*], George V. Iordache[*], Horacio González–Vélez[§]
[‡]Faculty of Automatic Control and Computers, University POLITEHNICA of Bucharest, Romania. https://upb.ro
[§]Cloud Competency Centre, National College of Ireland, Ireland. https://ncirl.ie/
[†]Digital Technology Skills Ltd., Ireland. https://digitaltechnologyskills.ie/
[*]BEIA Cercetare S.R.L, Romania. https://beia-cercetare.ro/

***Abstract*—Over the past five years, different organisations have increasingly called for science to become more open and reproducible. They have endorsed a set of data-management principles known as the FAIR (Findable, Accessible, Interoperable, Reusable) principles. As such, there is a growing trend towards the open availability of research data, as researchers continue to enhance reproducibility by enabling sharing and opening of their findings and datasets. However, there is not yet a standardised way to openly enable access to datasets while keeping control of their final use, potentially obtaining benefits from their utilisation. This paper introduces SMARDY, an EU-funded project which is deploying a traceable FAIR-compliant open innovation marketplace for data. Its innovative method for data exchange consists of the use of blockchain for controlling access rights to data, with data models able to grant access according to policies completely kept under the control of the data owner/producer. We also describe how SMARDY employs dimensionality reduction techniques to automatically generate FAIR–compliant metadata, statistical fingerprinting to identify derivated datasets, and watermarking to help data owners trace the distribution of multiple copies of a dataset.**

***Index Terms*—open data, open science, FAIR, data exchange, automatic versioning, blockchain, data repositories.**

I. Introduction

Research impact has been traditionally assessed in terms of accountability, funding generation, and management indicators rather than on the availability to generate better methods and advance findings [1]. As a traditional factor to determine scientific progress, research impact relies on peer review and bibliometric indicators rather than on the potential of experimental research results. Paradoxically, even when results show a meaningful and important avenue to further advance knowledge, their associated datasets are not always available to be analysed again.

Broadly defined as non-privacy-restricted and non-confidential data which is produced with public money and is made available without any restrictions on its usage or distribution, *Open Data* has revolutionised collaboration and reproducibility in science [2]. The European Commission estimated that its market size was to reach a value of 75.7 bn EUR by 2020, with over $100,000$ direct jobs and cost savings of 1.7 bn EUR for the European public sector [3]. The implications and benefits of sharing and openness are clear for researchers: they can catalyse new collaborations, increase confidence, and generate goodwill. Datasets are becoming easier to cite and bibliometric indicators enable researchers to get credit for their datasets. For public funding agencies, Open Data implies, in general, Open Science. Ergo additional discovery can move forward on top of already existing results, as well as double-funding ideas already funded can be avoided.

On the one hand, different organisations are actively supporting improved data access and have called for science to become more open by endorsing a set of data-management standards known as the FAIR (Findable, Accessible, Interoperable, Reusable) principles [4]. On the other hand, it has become increasingly important to justify the value generated from research datasets, whether from government agencies in the interest of transparency and accountability [5], or from commercial entities in search of ecosystems to foster innovation [6]. It has become clear that there are no standardised ways of tracking and measuring the tangible benefits from research data and that there are also significant issues associated with research data management such as copyright and ownership, data licensing, erroneous interpretation of data, and data security and privacy [7].

> While scientific discovery is arguably moving towards a greater data openness, there is no standardised way to keep data provenance, marshal FAIR guidelines, and track sources and ownership.

That is to say, new solutions for research data sharing require standard data and metadata representations, secure ways to keep data provenance and standard ways to automatically generate metadata fostering reusability. To overcome all these barriers, a holistic solution for data traceability should consider: who gets the data, if tampered versions of a dataset are being spread, and mechanisms for future references to data objects. This paper describes our proposal to achieve such a research data marketplace.

This paper introduces the European research project "Marketplace for technology transfer of R&I data, software and results" (Smardy), which is developing a data marketplace where academia, industry, and government, can exchange curated datasets, technology, and tools to foster economic and social development.

> The Smardy zero-trust decentralised research data platform fosters FAIR data sharing while keeping traceability and provenance to monitor data usage and, ultimately, technology transfer and intellectual property.

II. Related work

The idea of building research on top of already available findings, a.k.a. Secondary Data Analysis, has been construed as an empirical research approach to reapply the same basic principles as an initial study with primary (original) datasets and then furnish a new relevant protocol as any other research [8], [9]. Data can be directly sourced from multiple academic, government, inter-government, organisational and commercial repositories [10] and there has been a growing adoption of research software and data curation [11] to foster digital preservation of born-digital research artefacts. Coupled with the 2014 *Joint Declaration of Data Citation Principles* [12] to highlight the evidencing importance of data and its persistence, data accessibility, and the need to give scholarly credit to contributors, secondary data analysis has become an open-ended approach to systematic data-intensive research.

In addition to standard data and metadata representation, there is still the issue of tracking provenance. On this matter, we believe that the immutability offered by blockchain systems suits this goal. We argue that such immutability and optionally-disclosed ownership make blockchain suitable to enable openness of data while keeping track of provenance. While blockchain systems for open data were initially circumscribed to rewards [13] or bitcoin analytics [14], more modern instances have started to look at the provenance monitoring [15], publication of linked open data [16], and data sharing on plasma technologies [17], but they have not considered a complete FAIR-compliant framework to enable the creation of value with persistent data ownership, cataloguing, and automatic generation of metadata.

Finally, it is noted that Smardy deploys an end-to-end zero-trust architecture [18] comprising identity and access management including institutional credentials, well-defined operations and endpoints using data management standards, distributed hosting environments, and cloud-enabled interconnecting infrastructure.

A. Research niche

Smardy nurtures a traceable open innovation environment by making use of a fully decentralised system for controlling access rights and publishing data catalogues. As a zero-trust architecture, Smardy aims to prevent data breaches and limit internal lateral movement by using statistical and locality-sensitive mechanisms to avoid data leakage of copies and tampered versions. Such innovative combination of techniques not only supports all properties needed to share the data in secure conditions, but also prevents data tampering in a systematic way.

The Smardy platform enables the upload, fingerprinting and indexing of research-generated datasets as open data. Open data has the purpose of implementing the functionalities that discourage users afraid of intellectual theft. Our initial proof of concept entails the use of time-series data.

Additionally, the final system should grant access according to policies completely kept under the control of the data owner/producer with specific emphasis on FAIR principles. From a technical perspective, FAIR plays an essential role in the objectives of open science to improve and accelerate scientific research to increase the engagement of society and to contribute significantly to economic growth. However, despite the existence of open access repositories to share data and promote its re-usability, *there is not yet a clear open mechanism to trace how data is further used by others and, possibly, monetise their use*. This is another research question bringing novelty and innovation to our project.

III. Proposed solution

Finding proper inputs is essential for any secondary research project, and we have built the entire platform around the **Data Repository**, which indexes all the datasets in the Smardy platform. We have initially chosen to extend the Comprehensive Knowledge Archive Network (CKAN) [1], an open-source data management system for powering data hubs and data portals. CKAN is an interoperable platform that enables the open world to import the datasets into similar data repositories. It also enables datasets to be searchable not only through the local instance but also through other more significant and more visible (national, regional or even worldwide) data repositories.

To have Single-Sign-On (SSO) across the Smardy platform, we have employed a subset of trusted Identity Providers (IdPs) to identify the academic/research users. Such IdPs also fit the minimal requirements towards a Zero-Trust environment by: i) strongly verifying the user's identity; ii) including device validation (many times, the user can make SSO login only through trusted devices); and, iii) offering proper affiliation/status information to the service provider (the platform that the user logs into) to provide a minimal level of privilege. EduGAIN[2] is a pan-European interfederation service sponsored by GEANT, which includes IdPs from more than $4,500$ research and education centres.

Fig. 1 presents the high-level Smardy architecture. Smardy enables two options based on the type of a dataset:

1) Public data can be searched for and downloaded without any restriction;
2) Protected and private datasets are also indexed in the data catalogue; the user can search them based on their metadata, but these datasets are not stored in the data

[1] https://ckan.org

[2] https://edugain.org/

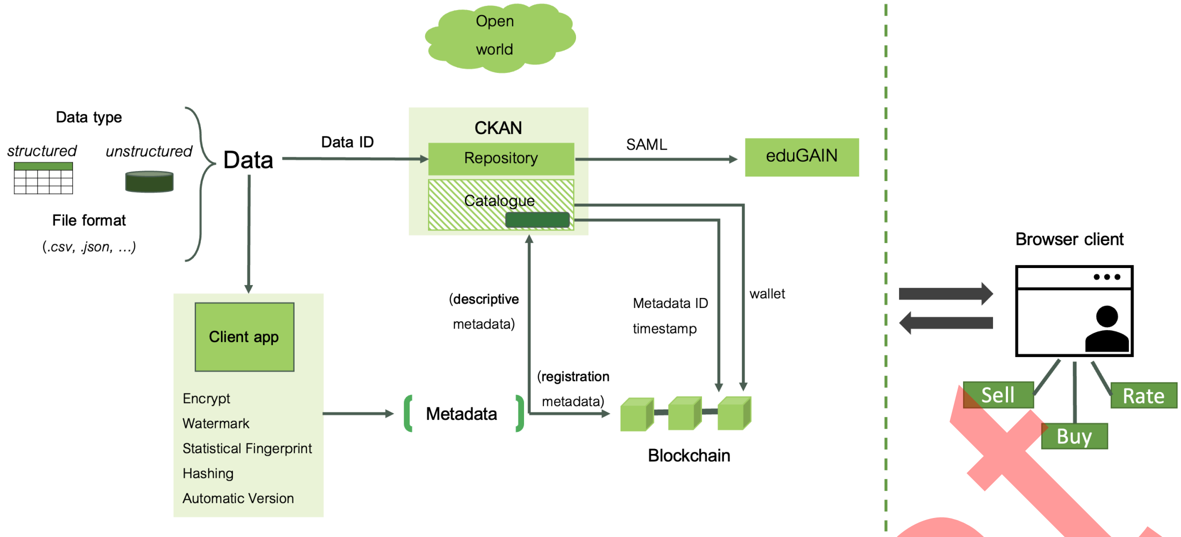

Fig. 1: SMARDY Architecture. This figure illustrates an instance generated by the upload of a dataset to the platform.

repository. Protected datasets get published to the IPFS in an encrypted version, and the potential buyer obtains the decryption key only through a blockchain transaction approved by the data owner. For this kind of dataset, instead of a direct file download, the data repository will provide a link to start the negotiation with the seller and—if they agree on an exchange—to initiate the blockchain transaction to buy the dataset.

In our system, the life-cycle of a protected dataset consists of the following steps:

- The data owner publishes the metadata about the offered dataset and prepares an encrypted version of it;
- The data buyer adds a data request in the blockchain;
- The seller (after receiving the desired compensation) agrees on releasing the data and provides the seller the data (as well as the decryption key).

1) Data ownership protection: In addition to being an open gate for searching valuable datasets and an innovative zero-trust marketplace for datasets, an important goal of the **SMARDY platform is helping the data owners to protect their intellectual property** and not just as a centralised platform witnessing the correct flow of ownership, but **as a decentralised platform offering non-disputable proofs of datasets being processed in the platform**. In order to achieve this, the metadatas (extended dataset description that helps to identify the datasets) are published in the blockchain ledger.

An example is when somebody steals a dataset in its identical to the original form and pretends to be the creator of that dataset as per a specified date, the truthful owner—who previously published the metadata of the files to the blockchain—can easily prove that he/she owned the file at the moment of metadata publication if a hash of the dataset is included in the dataset.

Considering the multiple forms of dishonest use of a dataset, we extend the simple exact match of 2 copies of a file with two advanced mechanisms:

- Using **statistical fingerprinting** of the datasets, that allows identifying derivative products (which might appear when a malicious user pretends to be the creator of a dataset obtained by altering an original dataset). We employ multiple fingerprinting methods, starting from those meant for the evaluation of a knowledgeable person (which might, for example, compare two data series based on their similarity of the median, average, number of points and meaning) up to those (such as TLSH, a locality sensitive hash) generating special hashes which programmatically denote the similarity of two pieces of information;
- Using **watermarking**, we aim to help data owners to protect their rights better when distributing multiple copies of a dataset to multiple destinations. In this case, based on a watermark extraction process, the data owner can uniquely identify which of the copies was involved in a dishonest use of data.

2) Lack of centralised authority controlling the data: One of the biggest concerns when talking about the online processing of data with limited audience is trusting the platform handling the operations. In SMARDY, we solve this problem due to the fact that protected datasets are never processed online in the non-encrypted form.

All the operations on the non-encrypted data are handled by the client application, which runs on the data owner's computer. The protected data only gets uploaded/transferred to another party only after it is securely encrypted (based on a key which is only known by the data owner).

The primary high-level function of the client application is to allow the data owner/data-seller to add datasets to the platform. This software component runs directly on the computer of the data owner to ensure that we maintain our zero-trust environment, in which the data seller does not have to trust the platform itself and maintains 100% control over the owned data.

The client application allows the data owner to extract the proper metadata for protected/private datasets without uploading any unencrypted data. This process also includes extracting statistical features and similarity hashes that anybody can use to verify against potential similar and derivative datasets. The data owner will upload the resulting metadata (consisting of all secret-preserving information about the dataset) to the data repository and the IPFS/blockchain to enable the finding of the dataset and offer an indisputable set of truths about the newly added dataset.

For protected datasets, the client application will also create an encrypted version that will be published in the blockchain to be released based on a potential transaction.

If the user wants to sell a watermarked dataset, one or more watermarked and encrypted versions will be generated. Such an approach is needed since watermarking is a process that should be done on the unencrypted version of the file, and, furthermore, the aim is to keep the uploading user in full control of their datasets (which are thus encrypted after watermarking).

3) Buy and rate: In addition to providing a safe environment to the academics/data owners willing to add datasets to the platform, Smardy provides advantages to a buyer in case of a transaction of a protected dataset. The main advantages are the ease of use and the rating system.

In our proposed solution, every action guarantees that the buyer should be able to provide a rating about the acquired product. Moreover, the *Know Your Customer* (KYC) approach based on trusted identity providers creates a good place for academics to meet the need for protected data ownership with the ease of involving in secondary research based on existing input data.

IV. Current State and Results

The proposed solution is part of the ongoing Smardy research project. In this section, we present the current status of the prototype development as well as some interesting results regarding the chosen solutions.

A. Dimensionality Reduction

Data provenance and lineage play a key role in fostering reusability. In this context, we propose a novel approach for automatic data versioning with the goal to automatically generate FAIR-compliant provenance metadata [19]. We systematically detect and measure changes in datasets by using dimensionality reduction techniques. A dimensionality reduction technique provides a low-dimensional model of the original dataset while still preserving the main characteristics of the original data. Our approach employs parameters from two dimensionality reduction approaches, namely *Principal Component Analysis* (PCA) and *Autoencoders*, to quantify the similarity between datasets versions. These parameters can then be included in the metadata to enable data versioning and data comparison. Fig. 2 summarises our approach.

We empirically evaluated our approach using time series data in three common scenarios of data versioning, namely i) data (cells) with missing values, ii) data with the row-wise transformation of values (values expressed as percentages) and data with column-wise transformation (values expressed on a logarithmic scale), and iii) sample size reduction by sub-setting rows. Initial results presented in [19] show that our proposed approach successfully detects different versions of a dataset, for up to 60% of cell changes, the deletion of up to 60% of rows, and column-wise transformation. However, there are minimal similarities detected for row-wise transformations. Also, the results show that, in general, the PCA-based similarity metrics provide a more robust indicator when comparing the original data with a changed data version compared to the Autoencoder-based similarity metrics.

B. Data Fingerprinting and Watermarking

Due to the variety of file types and their size, Data Fingerprinting uses a similarity algorithm to provide resistance to randomised attacks and low false positive detection rates, specifically *Locality Sensitive Hashing Algorithm (TLSH)* [20].

Compared to the rest of *LSH* algorithms, *TLSH* outperforms them, especially when missed detection is of concern.

One of the main strengths of this algorithm is the usage of a distance score, where a score of 0 represents that the files are identical and scores above that represent the difference between the files. The rest of the schemes, like `ssdeep` and `sdhash`, provide a similarity score that ranges from 0 to 100. Once a score is reduced to zero, then the schemes cannot adjust their threshold any further. An open ended distance criteria makes the job of changing the file more difficult [21]. Also, when mutations of the entire file happen, the *TLSH* algorithm still provides a low distance score, as compared to `ssdeep` and `sdhash` algorithms, for example, which return a similarity score close to 0 after a small number of alterations.

Watermarking uses broadly used file formats, such as *CSV* and *JSON*. Starting from this point, we decided to use a semi-structured data protection scheme based on robust watermarking [22]. Since the main scope of this technique is to keep the data intact, this schema keeps the distortion of data relatively small. It also allows us to select the appropriate fields of the *CSV* or *JSON* files to which we want to apply the watermark. This aspect of the schema is very important, since parts of the original file should remain intact.

By using watermarking, we aim to help data owners to better protect their rights in case of distributing multiple copies of a dataset to multiple destinations. In this case, different watermarks will be generated for the same dataset. Having a different watermark for each distributed copy of the dataset and an entry in blockchain with the logged transactions, the rightful owner could easily determine who was involved in a dishonest use of data. In terms of watermarking extraction, the algorithm presents a high ratio of successful extraction under different attacks, such as deletion, insertion and modification of up to 50%, on both *CSV* and *JSON* files. The extraction

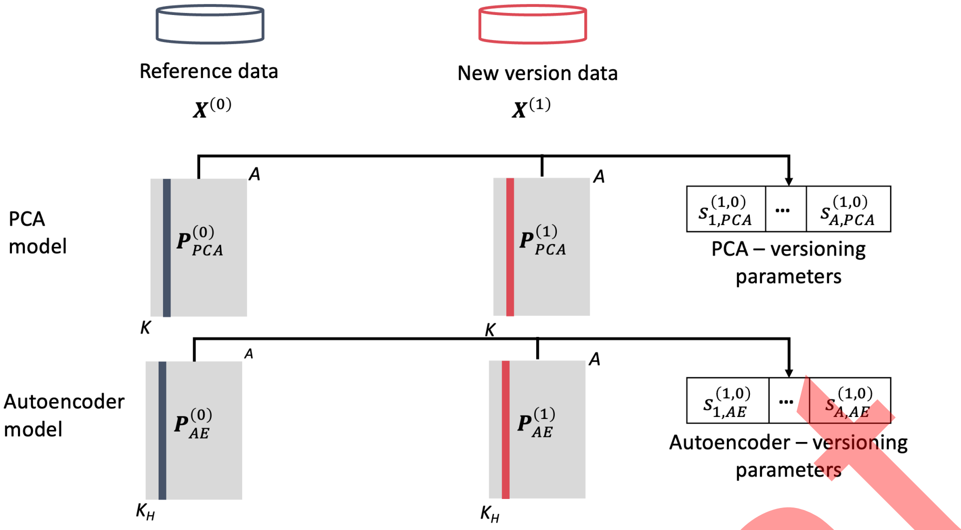
Reference data
$\boldsymbol{X}^{(0)}$
New version data
$\boldsymbol{X}^{(1)}$
PCA model
$\boldsymbol{P}_{PCA}^{(0)}$
A
K
$\boldsymbol{P}_{PCA}^{(1)}$
A
K
$s_{1,PCA}^{(1,0)}$ ... $s_{A,PCA}^{(1,0)}$
PCA – versioning parameters
Autoencoder model
$\boldsymbol{P}_{AE}^{(0)}$
A
K_H
$\boldsymbol{P}_{AE}^{(1)}$
A
K_H
$s_{1,AE}^{(1,0)}$ ... $s_{A,AE}^{(1,0)}$
Autoencoder – versioning parameters

Fig. 2: Automatic versioning: calculation of versioning parameters by comparing the PCA model (top) or the Autoencoder model (bottom) of the reference data ($\mathbf{X}^{(0)}$) and a newer version of data ($\mathbf{X}^{(1)}$). These parameters are then included in the metadata to enable data versioning and comparison. [source: [19]].

is the most successful on larger files, which is a plus for our proposed solution.

We present some initial results regarding just fingerprinting and watermarking. Both the techniques, separately, can be used with the intention of protecting data. But using them together will represent a powerful tool since they are not dependent on each other.

The client application, which will run these algorithms, must first embed the watermark in the dataset and then calculate the *Locality Sensitive Hash*. Having this in mind, we will present the result of altering a dataset after we embedded the watermark and calculated the hash.

The dataset we use is a large and simple JSON file with data extracted from a weather station, with values for temperature, atmospheric pressure, wind, humidity and other parameters. The file is 22MB in size and has approximately $170,000$ entries. The dataset describes the name and type of the sensor, the recording date and the value.

The properties of the dataset that is to be used in the empirical evaluation are the ones of big datasets (the **volume** depends on what measured parameters will be used from the historical data sets, for example in future experiments we will use volumes of data in the range of tens or hundreds of MBs, because weather related datasets are big in size; the **variety** of the data is given by the diversity of the measured weather parameters considered; and the **velocity** of the data, the measurements, used to gain value, will depend on how fast the measurements will be performed as time series). One of the uses of weather related data sets, which is big in size, is the process of weather forecasting.

To keep the example as simple as possible, the watermark that will be embedded in the file is `smardy`. The next step is to calculate the fingerprint of the file using the TLSH algorithm. The value generated is:

TABLE I: Alterations to the dataset.

Operation	Distance	Watermark
Delete 10 entries	1	Yes
Move 22 entries	1	Yes
Add 2400 entries	1	Yes
Replace ‘2022‘ with ‘2023‘ in the entire dataset	2	Yes
Delete 6300 entries	2	Yes
Replace ‘0.‘ with ‘3.‘ in the entire dataset	2	Yes
Delete 1550 entries	2	Yes
Move 4000 entries	2	Yes
Add 6500 entries	2	Yes
Rename fields	72	Yes
Replace 6 with 9 in the entire dataset	8	No
Rename fields and add random values	121	No
Delete 19000 entries	4	No

```
T12C37C36BF8601C7F0E3852
B265795786F3A4231F914D4A
823A3C8D4C3FB2D29B587D96
```

The sensitive hash will always have the same dimension, regardless of the file size. Having these steps complete, we will start altering the file and see how the proposed techniques behave. The operations performed on the file, as well as the results, are presented in Table I. After a short analysis of Table I we notice that there are cases in which the watermark could not be extracted, or the value of the distance is high.

There is a single case in which the values had spiked, namely when we renamed all the fields of the entities. This happened because *LSH* algorithms are text-based, and altering the field names of the entire file caused a huge difference from the previous one. However, these values should not worry us since the *TLSH* distance scores go up to 300. In the case where the distance is **72** we were able to extract the watermark and the score may represent the fact that the datasets are 74% similar. Having these values, the data owner can justify that the data belongs to them. In the other case, where the distance value is **121** the watermark could not be extracted and the value represents the fact that the files are around 50% similar. But adding random values in the file does not represent a way in which the dataset will remain usable.

It can be noted that in the last 3 cases, the watermark could not be extracted. It is worth mentioning that the watermark technique depends on the numerical values and especially on the decimals of the values. Deleting a large set of entries or altering numeric values will eventually make extracting the watermark impossible. However, in 2 out of 3 cases, the distance score is low, and the owner can easily prove that they owns that dataset.

Of course, there are a lot of cases and tests we can experiment with on different datasets with different operations. Nevertheless, even though those procedures are not perfect, we have seen that those changes surpassing both fingerprinting and watermarking are usually bound to highly-altered and unusable derived datasets.

C. Blockchain

To fulfil the goal of *zero-trust*, we must provide a mechanism through which the user can prove the authenticity of the data and also allow the user to share the data securely, without the mediation of a third-party. These can be acquired through blockchain technology. Having stored in the blockchain, in a secured manner, a copy of the dataset with its watermark and fingerprint, the user will be able to manually select the group of interest with whom he can share the data and later prove independently that he/she is the data owner by using all the techniques described in this section.

We propose the use of the *Ethereum* blockchain due to its widespread use and the simple and fast way to implement and deploy *smart contracts*. Ethereum uses *Solidity* [23] as programming language for the deployment of *smart contracts*. Each dataset will be shared as a *non-fungible token* (NFT). The proposed schema of the entity saved through the *smart contract* is described in Table II.

D. Client Application

The client application is mandatory for implementing the described techniques, while ensuring that the platform remains as transparent as possible. Due to the diversified operating systems and platforms, we will use a cross-platform framework such as `.NET MAUI` or `Ionic`. It is still to be determined which framework will be able to take advantage of all the schemas presented since they are written in different programming languages and use different technologies.

The client application enables to purchase protected datasets in a multi-step process (also depicted in Fig. 3) as follows:

TABLE II: Entity created by the Smart Contract.

Field name	**Field type**	**Description**
itemId	uint	unique id for the item
productId	uint	identical for items representing the same product, but with a different watermark
buyerId	uint	buyer unique id
sellerId	uint	seller unique id
nft	IERC721	NFT smart contract instance
isForSale	bool	determines if the dataset is for sale or if it is uploaded just for proof of publication
isSold	bool	determines if the item is sold or not
price	uint	the price of information if it's for sale
ipfsLink	string	the location of the information if it is for sale - the watermarked information is stored in an encrypted manner
fileHash	string	the hash of the information stored in IPFS
sellerAddress	address	instance of sellers address
sellerRate	uint	after the information is sold, the buyer could give a rate to the seller
productRate	uint	after the information is sold, the buyer could give a rate to the product
encryptedKey	bytes32	used to decrypt the file stored in IPFS; the key is saved in an encrypted manner, using a session key

- **The buyer** finds the dataset in the data repository (CKAN) and gets redirected to the blockchain-based application;
- **The buyer** calls a smart contact to create a data transfer request;
- **The data owner** approves the transfer and:
  - The decryption key gets released to the seller;
  - The sell gets logged into the blockchain.
- **The buyer** is entitled to offer a rating for the acquired dataset.

As required by the zero-trust standards [24]: *(i)* the purchased item is not disclosed to any unintended 3rd party (including the marketplace); *(ii)* since data is encrypted, no authority can decide to disclose protected datasets without the approval of the data owner; *(iii)* the proof of a transaction is permanent and it cannot be hidden; and *(iv)* once the transaction was registered, nobody can prevent the data buyer from publishing a rating. Moreover, the content of the reviews can not be changed or deleted.

V. Conclusions

This paper presents the design and evaluation of a zero-trust marketplace for research data. We have presented our initial results on how Smardy employs dimensionality reduction techniques to automatically generate FAIR–compliant metadata, statistical fingerprinting to identify derivated datasets, and watermarking to help data owners trace the distribution of multiple copies of a dataset. Such techniques are arguably enabling data openness while preserving provenance and ownership.

An important contribution of our proposed solution consists of the use of blockchain for controlling access rights to data, with data models able to grant access according to policies completely kept under the control of the data owner/producer. Furthermore, our assumptions of no implicit trust zones or

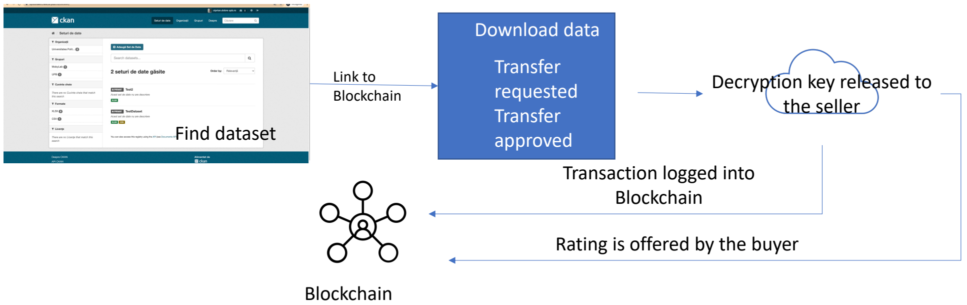

Fig. 3: Buy-release-rate process for protected datasets.

resources (datasets), geographically-distributed devices, no default institutional-owned infrastructure, and consistent workflows have allowed the SMARDY approach to enable some zero-trust fundamentals.

Our initial results indicate that the SMARDY system is increasingly enabling the creation of value by allowing data owners to monitor the usage of datasets while improving Findability, Accessibility, Interoperability and Reusability of datasets, i.e. making datasets more FAIR–compliant.

ACKNOWLEDGEMENT

This work has been developed as part of the European research project "SMARDY: Marketplace for Technology Transfer of Research Data, Software, and Results" (2021–2024) funded by the European Eureka Network programme through i) Romania's Ministry of Research, Innovation, and Digitalisation CNCS/CCDI-UEFISCDI (Ref#: PN-III-P3-3.5-EUK-2019-0241) PNCDI III; and, ii) Ireland's International Research Fund of Enterprise Ireland (Ref#: IR20210058). The research was also supported by the project DECIP: "Dezvoltarea Capacitatii Institutionale a Universitatii POLITEHNICA din Bucuresti", under contract no. PFE57/2022.

REFERENCES

[1] T. Penfield, M. J. Baker, R. Scoble, and M. C. Wykes, "Assessment, evaluations, and definitions of research impact: A review," *Research Evaluation*, vol. 23, pp. 21–32, 10 2013.

[2] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, adoption barriers and myths of open data and open government," *Information Systems Management*, vol. 29, no. 4, pp. 258–268, 2012.

[3] W. Carrara, W. S. Chan, S. Fischer, and E. van Steenbergen, "Creating value through open data: Study on the impact of re-use of public data resources," Tech. Rep. SMART: 2014-1072, European Commission, Directorate General for Communications Networks, Content and Technology, 2015. ISBN: 978-92-79-52791-3.

[4] M. Wilkinson, M. Dumontier, I. Aalbersberg, *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, pp. 160018:1–9, 2016.

[5] A. Zuiderwijk, R. Shinde, and M. Janssen, "Investigating the attainment of open government data objectives: is there a mismatch between objectives and results?," *International Review of Administrative Sciences*, vol. 85, no. 4, pp. 645–672, 2019.

[6] A. Zuiderwijk, M. Janssen, and C. Davis, "Innovation with open data: Essential elements of open data ecosystems," *Information Polity*, vol. 19, pp. 17–33, 1 2014.

[7] D. Patel, "Research data management: a conceptual framework," *Library Review*, vol. 65, no. 4–5, pp. 226–241, 2016.

[8] M. P. Johnston, "Secondary data analysis: A method of which the time has come," *Qualitative and Quantitative Methods in Libraries*, vol. 3, no. 3, pp. 619–626, 2014.

[9] L. Rodriguez, "An interdisciplinary approach to secondary qualitative data analysis: what, why and how," in *Handbook of Qualitative Research Methodologies in Workplace Contexts* (J. Crossman and S. Bordia, eds.), ch. 9, pp. 133–156, Cheltenham: Edward Elgar Publishing, 2021.

[10] E. Manu, J. Akotia, S. Sarhan, and A.-M. Mahamadu, "Identifying and sourcing data for secondary research," in *Secondary Research Methods in the Built Environment* (E. Manu and J. Akotia, eds.), ch. 2, pp. 16–25, London: Routledge, 2021.

[11] A. Chassanoff and M. Altman, "Curation as "interoperability with the future": Preserving scholarly research software in academic libraries," *Journal of the Association for Information Science and Technology*, vol. 71, no. 3, pp. 325–337, 2020.

[12] M. Altman, C. Borgman, M. Crosas, and M. Matone, "An introduction to the joint principles for data citation," *Bulletin of the Association for Information Science and Technology*, vol. 41, no. 3, pp. 43–45, 2015.

[13] S. Lynch, "OpenLitterMap.com – Open Data on Plastic Pollution with Blockchain Rewards (Littercoin)," *Open Geospatial Data, Software and Standards*, vol. 3, no. 6, pp. 1–10, 2018.

[14] D. McGinn, D. McIlwraith, and Y. Guo, "Towards open data blockchain analytics: a bitcoin perspective," *Royal Society Open Science*, vol. 5, no. 8, p. 180298, 2018.

[15] T. K. Dang and T. D. Anh, "A pragmatic blockchain based solution for managing provenance and characteristics in the open data context," in *Future Data and Security Engineering*, vol. 12466 of *LNCS*, (Vietnam), pp. 221–242, Springer, Nov. 2020.

[16] F. Kirstein and M. Hauswirth, "Blockchain for trustworthy publication and integration of Linked Open Data," in *11th Knowledge Capture Conference*, K-CAP '21, (Virtual Event, USA), p. 269–272, ACM, Dec. 2021.

[17] S. Tschirner, M. Röper, K. Zeuch, *et al.*, "Fostering open data using blockchain technology," in *DIONE 2021: Data and Information in Online Environments*, vol. 378 of *LNICST*, (Virtual Event), pp. 209–228, Springer, Mar. 2021.

[18] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero Trust Architecture," Special Publication NIST SP 800-207, National Institute of Standards and Technology, Gaithersburg, Aug. 2020. Online at: https://doi.org/10.6028/NIST.SP.800-207.

[19] A. González-Cebrián, L. A. McGuinness, M. Bradford, A. E. Chis, and H. González-Vélez, "Automatic versioning of time series datasets: a FAIR algorithmic approach," in *18th IEEE International Conference on eScience, eScience 2022*, (Salt Lake City), pp. 204–213, Oct. 2022.

[20] J. Oliver, C. Cheng, and Y. Chen, "TLSH–a locality sensitive hash," in *2013 Fourth Cybercrime and Trustworthy Computing Workshop*, (Sydney), pp. 7–13, IEEE, Nov. 2013.

[21] J. Oliver, S. Forman, and C. Cheng, "Using randomization to attack similarity digests," in *International Conference on Applications and Techniques in Information Security: ATIS 2014*, vol. 490 of *Communications in Computer and Information Science*, (Melbourne), pp. 199–210, Springer, 2014.

[22] J. He, Q. Ying, Z. Qian, G. Feng, and X. Zhang, "Semi-structured data protection scheme based on robust watermarking," *EURASIP Journal on Image and Video Processing*, vol. 2020, no. 1, pp. 1–10, 2020.

[23] M. Mukhopadhyay, *Ethereum Smart Contract Development: Build blockchain-based decentralized applications using solidity*. Packt Publishing Ltd, 2018.

[24] M. Shore, S. Zeadally, and A. Keshariya, "Zero trust: The what, how, why, and when," *Computer*, vol. 54, no. 11, pp. 26–35, 2021.