TECHNICAL UNIVERSITY OF CLUJ-NAPOCA

ACTA TECHNICA NAPOCENSIS

Series: Applied Mathematics, Mechanics, and Engineering
Vol. 67, Issue Special III, Jully, 2024

RESEARCH DATA INTEGRATION MODEL IN WEB 2.0 TO WEB3 TRANSITION

Nicolaie CIUBOTARU, Saltanat MEIRAMOVA, Angela REPANOVICI

Abstract: *Europe is traversing an intense period of transformation under the «Digital Decade» policy, being sustained through direct cooperation between the European Commission and the Member States. One of the main goals is to achieve a «digital transformation of businesses» with a specific objective concerning «small businesses and industry have access to data» where «innovative infrastructures converge to work together». This effort will be guided through a set of policies generically called the Digital Compass, but the concrete work will be done through multi-country large scale projects investing in areas like high-performance computing, a common data infrastructure, and blockchain technology among many others. In this paper we look into the transitioning period from Web2 technologies to Web3 technologies where some opportunities are already at hand in the field of research data, scholarly communication, and digital cultural heritage digital objects management. We place our investigation in the larger context of the activities dedicated to the European Year of Skills inviting all effort to upskilling needed for the digital transition. In this context we have analyzed the existing practices and we arrived in a possible transitional model for the data and metadata of research outputs without limiting the scope to it. New opportunities are emerging with the rise of Web3 and the distributed ledgers (blockchain), and these are backed in a proposed integration model.*

Key words: *research data, metadata, data management, web3, transition model, ipfs, blockchain, reactive research digital objects.*

1. INTRODUCTION

Understanding the perspective on the transition period from individual digital repositories based on the implementation of Internet and WWW technologies, to the offer of decentralization that Internet and Web3 technologies offer, it was necessary to investigate current solutions and models.

The first iteration of the World Wide Web was marked by the emergence of websites hosted by servers connected through the Internet. The pages presented by them were static, an aspect that led to the name read-only Web. Starting in 2004, web pages received the contribution of server technologies serving aggregated content on demand, and thus the read-write Web emerged generically being called Web 2.0. With a consolidation of decentralized technologies (which do not necessarily need a web server), virtual and augmented reality and IoT (Internet of Things), we can consider that we have entered the Web3 [1] era, which was introduced by the emergence of blockchain with its most famous application, Bitcoin. The preparation for this reality, this new information space, was partly achieved through the Semantic Web that was supposed to be, as the famous Tim-Berners Lee said, Web 3.0.

Because most often in the last ten years Restful APIs are responsible for data access/dissemination, they have been considered as an integral part of any implementation for any system that manages research data and metadata.

The main responsibility of APIs is to provide structured access according to the categories and typologies of data of interest, but the most important aspect is the standardization of this access. This study investigates a new possible model proposed below that is rooted in the

technologies offered by Web 3, so that API transactions use data stored in graphs that are hosted by distributed solutions and not in silos considering the practices mentioned in the paper.

2. METHODOLOGY

This study proposes a new possible model that has its root in a bigger study concerning the evaluation of systems for creating digital repositories with a focus on the components that create the means of data communication through APIs (Application Programming Interfaces). Existing APIs are an integral part of the big systems, sophisticated software implementations that are managing research data and metadata obtained out of the current practice.

The data is produced by the memory institutions (libraries, archives, museums), and international scientific information resource providers. The analysis of the APIs took into consideration 57 services offering the means to access a variate typology of data in various formats.

One issue that came to light was the diversity of the identifiers to the digital entities exposed through their endpoints. The metadata used to describe digital objects also exposes one researcher to various schemas, and often arrangements for serialization of the responses.

The dataset is accessible for consultation on Github at the https://github.com/kosson/apis-data-source where it will be kept alive and updated from now on.

One issue identified during the analysis was the multitude of identifiers (IRIs) used to make unique the entities describing the digital objects.

This observation spurred the interest in analysis and search for a possible true unique identifier unchained from the server location or the means of dereferencing (DNS). The search for an answer took our steps through the scientific literature provided by the following:

- International Conference on Theory and Practice of Digital Libraries (TPDL: [http://tpdl2023.dei.unipd.it/index.html](https://t.co/wqYlQZzGJl))
- International Journal on Digital Libraries (IJDL: [International Journal on Digital Libraries | Home] (https://www.springer.com/journal/799))
- Semantic Web in Libraries (SWIB: [SWIB23 Home](https://swib.org/)).

The bibliographic analysis revealed the existing solutions and revealed a clear overview of the existing technologies and how scholarly communication is benefiting from data exposure through APIs and digital repositories. What triggered this study was the need for the betterment of the technical ecosystem involving Web3 and blockchain where true unique identification and secure transactions should be taken into consideration.

3. INTEGRATING DIGITAL OBJECT DATA AND METADATA INTO REACTIVE ENTITIES

The European Data Strategy mentions that decentralized technologies offer the possibility to manage data flows based on free will and self-determination. The European Commission recognizes the potential of these technologies, which is part of the technical area they call Web3.

The organization of services that provide data/information using the World Wide Web presents a perspective of poles, of nodes that centralize data, information and services. We call this reality Web 2.0 or the World Wide Web in general. Web3 is a new organization of specialized information spaces using protocols focused on restoring trust in communicated data, based on cryptographic means.

The main attribute of Web3 networks and services is trust. It is derived from the fact that Web3 technologies allow for the verification of the status of transactions taking place on the network at any time. Thus, the areas of applicability are that of finance, but the most sensible for research data management would be that of identity management of the entities.

The premise is that no actor in the network can be trusted. Achieving the level of trust is done using blockchain. The distributed model is needed to update the blockchain with the current state or better said, with the last state of a

transaction. Until yesterday this role was fulfilled by a database.

Digital objects are the foundation of understanding the modern architectures of the components of the networks for the exploitation and distribution of scientific research results, as well as the objects belonging to the digital cultural heritage. Understanding them proves necessary to realize the context of the services and curatorial actions necessary for their digital management and preservation.

Recently, a new concept is making its way to the attention of researchers. It is RO-Crate (Research Object Crate) [2], an approach to simplify as much as possible the connection of research data with their metadata putting all the components of a research object in a so-called crate. Specification 1.1 already exists from 2020 with a first version issued in 2019.

We have the unanimously accepted working definition given by Khan and Wilensky in 2006: a digital object is a data structure whose principal components are digital material, or data, plus a unique identifier for this material, called a handle [3].

The DONA Foundation, created by the Corporation for National Research Initiatives (CNRI) in 2014, defines digital objects as a sequence of bits, or a set of sequences of bits, incorporating a work or portion of a work or other information in which a party has rights or interests, or in which there is value, each of the sequences being structured in a way that is interpretable by one or more of the computational facilities, and having as an essential element an associated unique persistent identifier. As an important landmark after the start of this initiative of the Dona Foundation, in 2015 the construction of the European Research Cloud also begins, focusing on digital research objects (FAIR Data Objects).

An important mention that the Dona Foundation makes is that a digital object should be understood as a digital entity ("ITU-T Recommendation database," 2013) according to ITU-T X.1255. This digital entity aligns with the concept of semantic artifact that the EOSC Interoperability Framework mentions in a report [4]. The text provides another interesting definition that somewhat simplifies the understanding of a digital object: objects that allow binding all critical information about any entity. [...] The act of defining a Digital Object is the act of defining a boundary around a set of data points.

Let's study a digital object from the perspective of the needs for it to be dynamic/reactive. It should have the ability to connect and natively respond to capitalization requirements in various scenarios. Such a digital object exhibits some basic characteristics:

- Is a single bytestream (a single file) or several in a compound that we generically call a resource because it is uniquely identifiable;
- Has metadata describing the resource to which can be added those that have special roles determined by the context of storage, exploitation, distribution and digital preservation;
- Has mechanisms to ensure a very high level of interaction. These mechanisms can be workflows (the workflows mentioned by the fair research objects) or even smart contracts. These mechanisms unitarily expose the resource to transform it from an inert object to a reactive one. At this level apis are the key to connectivity;
- A mechanism to manage versioning, integrity and authenticity. Such mechanisms are available today in working with git or ipfs (interplanetary file system).

We have seen that in the intended behaviors for digital repositories of the future [5] there is a requirement that updating a resource triggers notifications about what has changed. This need will be reflected in the future by developing some reactive mechanisms that will ensure this behavior. One of these mechanisms can be a smart contract in a certain blockchain. If the structure of the research object changes (versions), the smart contract will be notified and it will be executed triggering a cascade of events provided for each type of notification.

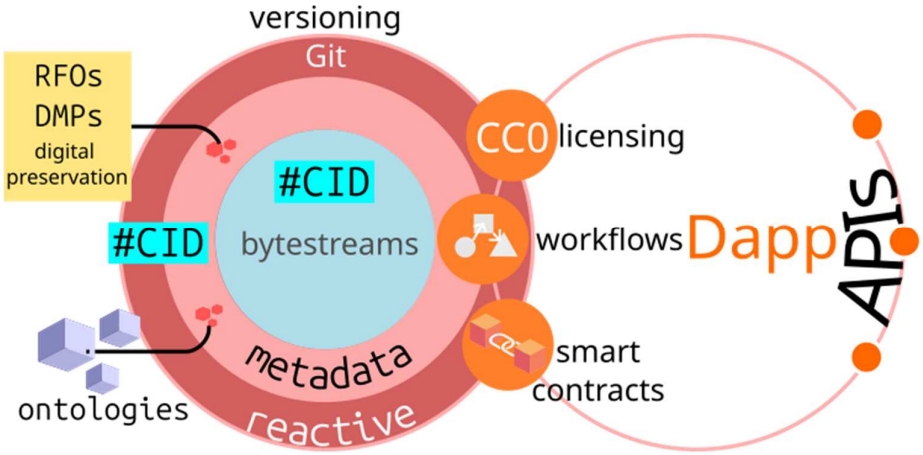

Fig. 1. Functional model of the digital object showing a reactive level (FAIR digital object or Cultural Heritage Object)

These notifications can be related to the modification of the research object, its introduction into a workflow or even its access through interaction with third-party services. What's more, whenever the research object is cited, the smart contract can receive a notification and thus execute it. A Smart contract is a software program that exists in a blockchain system with the role of keeping track of interactions/transactions made between actors.

In the not-too-distant future it is possible that scientific research articles will turn into the metadata of the research object that will be a reactive entity. This can be done by transforming the article, either by encoding possibly using TEI (Text Encoding Initiative), or by fully transforming it into a notebook (Jupyter Notebook). The closest model we've come across that comes close to the reactive entity we're describing is RO-Crate, which provides a sufficiently flexible data support structure but still passively waits to be integrated into a workflow.

Today, Web3 technologies are the ones that allow the transformation of inert and hard to kink digital objects of research or cultural heritage, whose level of interaction is achieved through the application that manages them, into real reactive digital objects.

Mainly, reactive digital objects in the field of research and cultural heritage will continue to use the mechanism of RESTful APIs to achieve standardized access, sufficiently richly documented and universally accepted as a model. APIs are the mechanisms most useful in making functional connections between the various data silos on the Web. In the 2020 report [6] there is a characterization of APIs that restores the understanding of the value of APIs from another perspective: APIs are technical contracts that can be seen as software products that have a value chain. The most important point that can be captured is that of the need to look at APIs as communication mechanisms between machines. But for this to be possible the APIs would need to be built and perform transactions based on real contracts.

From the experience of working with applications that carry out the implementation of standards for the distribution and interconnection of data, the appearance of horizontal assemblies of various APIs is observed. Their purpose is to create flows capable of managing digital objects in the sense of the OAIS (bitstreams) model. Communication is based on HTTP. What differs is the formula in which the various services are integrated. The prospect of creating a framework in which interoperability can be achieved presents current difficulties that must be overcome either through the unitary adoption of a set of best practices or through the federation of services. The EOSC Interoperability Framework report lists these difficulties.

IPLD (InterPlanetary Linked Data) [7] is a data model capable of acting as a bridge between various existing blockchain protocols, but most importantly, it provides the means to create identifiers for any type of data based on content, as opposed to location dependent URIs. IPFS

(Interplanetary File System) is the data layer, a Peer-to-Peer network architecture. Of particular interest is the ontologies needed to provide the digital object with the rich context and aspiration to become part of a knowledge base.

The Semantic Web promised a distributed architecture of services and data if "things" are identified with IRIs. This vision is limited in success and sustainability due to the nature of how the World Wide Web is built. The advantage that IPLD presents is the seamless integration of descriptive entities with the data/content they address.

Currently, Web3 technologies are starting to push the boundaries of decentralized financial applications (DeFI), providing solutions for multiple application areas. Below, in Figure 2, an overview of all actors in the transition from Web 2 to Web 3 is presented.

In the proposed general model, the premise is that digital objects (reactive entities) as well as metadata (identified generically by data structures) are considered objects that have their own CID (Content Identifier) generated. If the digital objects are part of an aggregate, the aggregate itself will have its own CID that can be generated by the algorithm's characteristic of Merkle tree data structures. All actors, data or metadata, in fact, are represented as Merkle trees, which gives flexibility when aggregation is needed, as well as preserving its own identity.

Thus, data management mechanisms will benefit from the exposure of increased granularity, but more than that, from new ways of distributed storage. It must not be forgotten for a moment that we are going beyond the limits of the WWW model, the current Web 2.0.

The aggregation of resources into complex Digital Objects such as RO-Crates, for example, transformed into reactive Digital Objects, will be reflected by a distinct graph of Interplanetary Linked Data type, which will allow to be exploited and interconnected with other similar graphs or which are currently used by implementations of Semantic Web technologies (knowledge graphs through the interconnection made through RDF – Resource Description Framework).

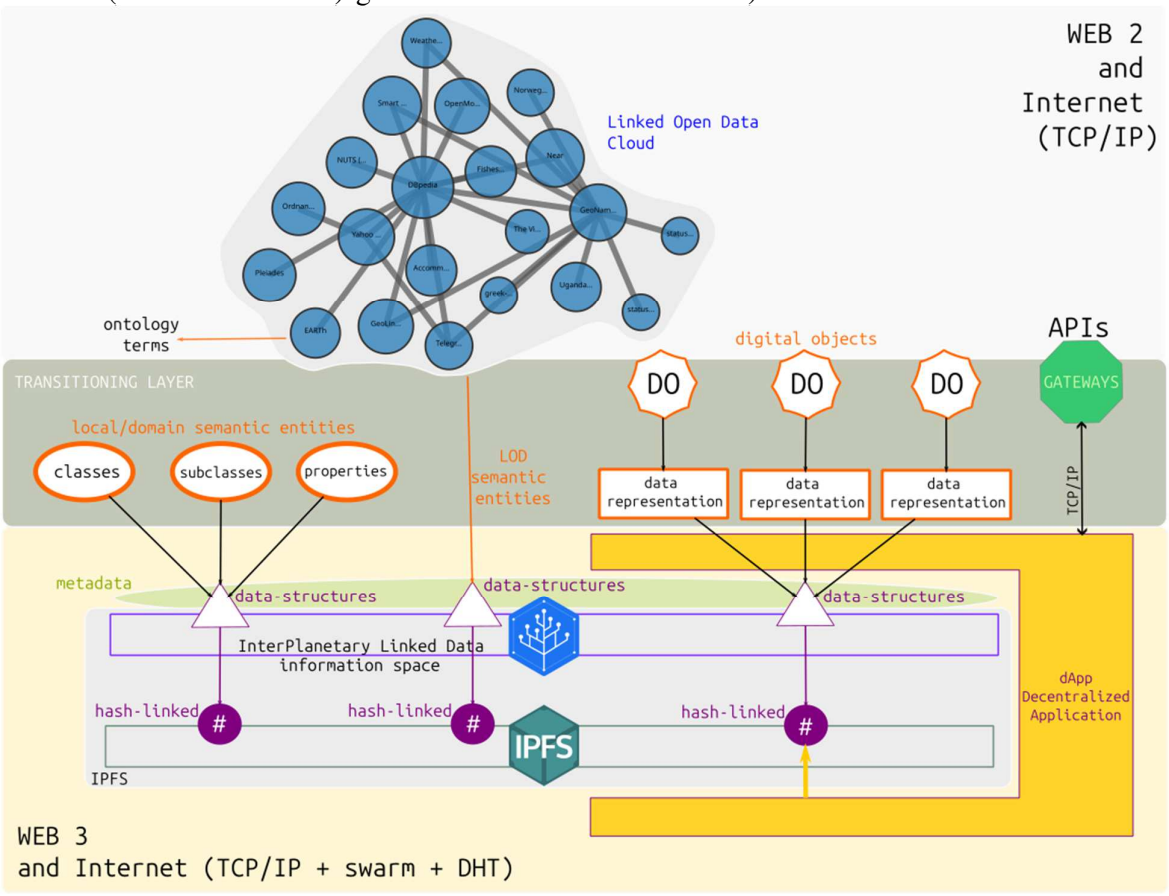

Fig. 1: General Model for Integrating Existing Semantic Entities with Web3's Distributed Technologies.

Beyond the digital objects themselves, even classes and subclasses of ontologies, and even ontologies themselves will be able to be represented by IPLD graphs. The vocabulary will be the first to benefit from a representation in the Interplanetary Linked Data information space. Each term will get its own CID, and the complex relationships they establish can be represented by IPLDs like thesauruses, for example. The most important gain is related to the fact that the unique identification will be given by a CID (Content Identifier), not by a link (IRI expressed as URL).

The advantage that the CID presents is that of undeniable uniqueness. Its value is an alphanumeric string resulting from the cryptographic processing of the contents of that file/resource/bytestream itself. While links/URIs uniquely identify a resource on a web server through a URL, there is no guarantee that that resource is unique. For example, there can be countless copies of the same EPUB or PDF file of a research article on countless servers, but whose identifier is a URL.

If the same resource resides on three servers, we could comfortably have three different URLs identifying the same resource/bytestream. We do not guarantee the uniqueness of the identifier. For this reason, the Digital Object Identifier was implemented as a solution to the need for a unique identification, but this mechanism is also tributary to Internet and WWW technologies, with all possible cases where a unique identification is not correctly achieved.

The CID solves this problem by analyzing the content, generating a result after cryptographic processing with specialized algorithms. The respective algorithms, whenever they have the same file as input, will produce the same unique identifier, the same CID.

The CID is not tied to any web location; however, its only dependency is the IPFS architecture. Management can be instrumented by creating specialized Dapps (Decentralized Applications). Dapps will play a role in orchestrating these resources. They will regulate access, determine authenticity and allow interaction with different digital objects based on sets of rules written in the smart contract. This is a very useful solution because if a Dapp proves ineffective, another can take its place (software development) and the data will no longer be part of the software solution, but will be part of IPFS, the management mechanism decentralized. This does not detract from proper administration by involving off-chain backups and enforcement of digital preservation regulations.

Because previously we designed the future research object as a reactive one, from the Web3 perspective, each of these can be managed through a smart contract. This can be done by integrating with blockchain implementations which is nothing but a distributed ledger (distributed ledger).

We can look beyond the transaction record mechanism because together with a unique identification that IPFS provides through IPLD, a context is created that gives independence and an increased level of interaction to digital objects. One of the problems that blockchain implementations have lies in the unique identification of the entities that carry out the transactions, and IPFS comes and elegantly solves this requirement by cryptographically issuing CIDs based on the content of the actors' resources.

The IPFS provides the truly unique identifier and the distributed ledger guarantees the immutability of transactions between identified entities. Another aspect that IPFS provides is related to the size of the digital object representation. Large digital objects cannot be stored in blockchain systems because the purpose for which they were created is different, thus being severely limited. Instead, by IPFS coupled with storage services like Filecoin, for example, this obstacle is no longer an issue.

At this point, because we have a concrete perspective of the fundamentals, we can imagine a scenario where a scientific research article is a smart contract, i.e. a reactive entity whose life cycle is regulated directly by source code. This entity can receive signals regarding certain metrics such as how many times it is cited or which works it cites. In the last scenario, a relationship between smart contracts is thus created, which can lead to tokenization (the

creation of a unit of quantifiable value) with the potential to reward the scientific contribution according to the novelty and value that the article presents. This data will be written in the blockchain and will be secure, tamper-proof with all the advantages of the cryptographic fingerprint it presents.

Another interaction may be between the research article and its accompanying data set. The data set can in turn be a reactive entity, that is, its life cycle can be regulated by the rules written in the smart contract. In a scenario of interaction of the scientific article with the dataset, we can very easily get very useful data on how the dataset has been used. If this is also a reactive entity, we can write very useful usage metrics into the blockchain. Moreover, as in the case of RO-Crates, we could introduce the rules of use in the allowed processing flows into the smart contract.

In the context of the proposed model, APIs become more important than ever because they are exercised whenever the software components involved are articulated, are composable. This flexibility can only be achieved by defining specific APIs.

Web APIs will play a special role in interconnecting Dapps. This can prove to be a real Achilles' heel for the proposed model if the data level is closely related to the business logic of the software solution. There is a danger of creating spaces separated by the degree of usability of a Dapp or a certain class of smart contracts. This is obvious if we think that smart contracts will be able to be created to manage classes of digital objects. Some will be able to manage the interaction with data sets, others with digital objects from the field of digital cultural heritage, others will have the role of collecting metrics, etc.

4. CONCLUSIONS

Semantic Web technologies have been heavily and unevenly assimilated over the past 10 years. Only at this moment when the question of rigors related to the need to ensure interoperability arises, semantic technologies prove their full usefulness. However, to create a sufficiently refined culture among those who will interact with such services that involve the understanding of semantic technologies, investments of time and above all in the training of all parties involved will be required.

It looks to the future to build a model of interactions between Dapps and the interaction of digital objects outside of Dapps. This is obvious because since they will start a life cycle in a distributed environment like Web3, mechanisms can be created to ensure an autonomous mode of interaction, beyond orchestration/administration through Dapps. Here we are envisaging fully automated workflows that don't need an administrative core, but a simple call from a human operator or not.

Currently, there is a need to create an enabling context for the training and retraining of all those who have active roles in the field of information sciences. The year 2023 is the European Year of Skills, this being one of the most important cardinal points of the Digital Decade. A worrying fact that the European Commission exposes in the motivation through the Digital economy and society index (DESI - Digital economy and society index) is that 4 out of 10 adults and one out of three people working in Europe do not have basic digital skills. This conclusion takes on an even more pronounced dimension in the context in which the Commission recommends more than once, as in the field of research, on the dimension of data management, that Member States invest in training programs specific to the needs that the digital objects require. From the managerial level to the one dedicated to digital preservation.

APIs will grow in visibility pushed by European policies seeking a level of interoperability in the data spaces they create (see Data Strategy). Now, APIs are the most targeted mechanisms for interconnecting systems, but also for exchanging data and metadata between silos (digital repositories). APIs are not tools to eliminate the shortcomings of the fragmented information space that the World Wide Web presents due to the server-client architecture on which it is based. Instead, it can create active bridges, standardized described services that can provide a level of security and uniformity.

5. REFERENCES

[1] *What is Web3 and why is it important?* (2023, September 14). [Blog]. Ethereum.Org. https://ethereum.org

[2] P. Sefton et al., *RO-Crate Metadata Specification 1.1.3,* Apr. 2023, doi: 10.5281/zenodo.7867028.

[3] R. E. Kahn and R. Wilensky, *A framework for distributed digital object services,* Int. J. Digit. Libr., vol. 6, no. 2, pp. 115–123, 2006, doi: 10.1007/s00799-005-0128-x.

[4] Directorate-General for Research and Innovation (European Commission) et al., *EOSC interoperability framework: report from the EOSC Executive Board Working Groups FAIR and Architecture*. LU: Publications Office of the European Union, 2021. Accessed: Sep. 17, 2023. [Online]. Available: https://data.europa.eu/doi/10.2777/620649

[5] E. Rodrigues et al., *Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group*, Nov. 2017, doi: 10.5281/zenodo.1215014.

[6] L. Vaccari et al., *Application Programming Interfaces in Governments: Why, what and how*, JRC Publications Repository, Sep. 15, 2020. https://publications.jrc.ec.europa.eu/repository/handle/JRC120429 (accessed Apr. 05, 2023).

[7] Protocol Labs, *IPLD - The data model of the content-addressable web*, IPLD. https://ipld.io/ (accessed Sep. 17, 2023).

Model de integrare a datelor de cercetare în contextul tranziției de la Web 2.0 la Web3

Europa traversează o perioadă intensă de transformare în cadrul politicii Deceniului Digital, fiind susținută prin cooperarea directă între Comisia Europeană și Statele membre. Unul dintre obiectivele principale este realizarea unei «transformări digitale a afacerilor» cu un obiectiv specific care privește «accesul la date a întreprinderilor mici și a industriei» în contextul căruia «infrastructurile inovatoare converg pentru a lucra împreună». Acest efort va fi ghidat printr-un set de politici numite generic Busola digitală, dar munca concretă va fi realizată prin proiecte la scară largă în mai multe țări care investesc în domenii precum calculul de înaltă performanță, o infrastructură comună de date și tehnologia blockchain printre altele. În această lucrare ne uităm la perioada de tranziție de la tehnologiile Web2 la tehnologiile Web3 în care unele oportunități sunt deja la îndemână privind datele de cercetare, comunicarea științifică și managementului obiectelor digitale specifice, dar și cele ale patrimoniului cultural digital. Ne plasăm investigația în contextul mai larg al activităților dedicate Anului european al competențelor, care invită la eforturi pentru îmbunătățirea competențelor necesare tranziției digitale. În acest context, am analizat practicile existente și am ajuns la un posibil model pentru datele și metadatele rezultatelor cercetării, fără a limita domeniul de aplicare doar la acesta. Noi oportunități apar odată cu creșterea Web3 și a blockchain-ului, acestea fiind susținute de un model de integrare pe care îl propunem.

Nicolaie CIUBOTARU, PhD, Transilvania University of Braşov, Mechatronics and Environment Department, Mechatronics and Environment Department, nicolaie.ciubotaru@unitbv.ro, 29 Eroilor Blv., Brasov, Romania.

Saltanat MEIRAMOVA, Dr. Saken Seifullin Kazakh Agrotechnical University, International Cooperation and Multilingual Education Development Centre, s.meiramova@kazatu.kz

Angela REPANOVICI, Professor, PhD. Eng., PhD Marketing, Transilvania University of Braşov, Faculty of Product Design and Environment, arepanovici@unitbv.ro, 29 Eroilor Blv., Braşov, Romania.