 DOI - https://doi.org/10.1145/3308560.3317076

# The FAIR TRADE Framework for Assessing Decentralised Data Solutions

John Domingue

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK, john.domingue@open.-ac.uk

Allan Third

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK, allan.third@open.ac.uk

Manoharan Ramachandran

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK, manoharan.ramachand-ran@open.ac.uk

## ABSTRACT

Decentralised data solutions bring their own sets of capabilities, requirements and issues not necessarily present in centralised solutions. In order to compare the properties of different approaches or tools for management of decentralised data, it is important to have a common evaluation framework. We present a set of dimensions relevant to data management in decentralised contexts and use them to define principles extending the FAIR framework, initially developed for open research data. By characterising a range of different data solutions or approaches by how TRusted, Autonomous, Distributed and dEcentralised, in addition to how Findable, Accessible, Interoperable and Reusable, they are, we show that our FAIR TRADE framework is useful for describing and evaluating the management of decentralised data solutions, and aim to contribute to the development of best practice in a developing field.

## CCS CONCEPTS

• Information systems~Distributed storage • Information systems~Resource Description Framework (RDF) • Security and privacy~Trust frameworks • Computing methodologies~Model verification and validation

## KEYWORDS

Distributed Data, Evaluation, Open Data, Semantic Blockchain

# 1 Introduction

We now live in a world where a small number of large technology companies hold a significant amount of personal data. For example, the Guardian journalist Dylan Curran (Curran 2018) found that Google had over 5.5GB of personal data on him including everywhere he had been, and all of his Google and YouTube searches (including deleted searches). His analogous figure for Facebook was over 600MB of data. As highlighted in a number of recent scandals in the press, such as those associated with the misuse of data by Facebook and the now defunct company Cambridge Analytica (Guardian 2015), over-centralisation can cause problems for individuals as well as negatively affect democracy and national culture. The current situation has resulted in, according to the Web's inventor, "producing—with no deliberate action of the people who designed the platform—a large-scale emergent phenomenon which is antihuman." (Lee 2018a). There is also now a growing recognition of the social problems caused when centralised data systems are impoverished with respect to disadvantaged citizens, who are in effect 'data poor'. A report by the US Government's Consumer Financial Protection Bureau (Bureau 2015) found that 26 million US consumers (11% of the adult population) are 'credit invisible' and a further 19 million (8.3% of the adult population)

had credit data that was unscorable. In the UK, a lack of citizenship data led to 63 British citizens being deported (Grierson 2018) and up to 57,000 UK citizens could lose rights to homes, jobs, social benefits, UK National Health Service treatment, or be threatened with deportation in the near future (Fact 2018) – a situation which has been blamed on the combination of poorly-managed centralized data and a lack of official recognition of citizen-held data.

One type of approach to resolving the above problem has been through decentralising the data. In particular, replacing central data controllers with either the user herself, or with peer communities. Giving users control over their data and how it is processed is seen by a growing number of researchers and developers as the best mechanism to empower users and give us the 'internet we want'. The best known of these is the distributed ledger (Walport 2015) an architecture for maintaining a peer network of verified data which underpins the Bitcoin cryptocurrency. Within the Linked Data (LD) arena, Solid (Lee 2018b) provides users with their own data pods, and Linked Data Fragments (LDF) (Verborgh et al. 2016) enables data processing to be split between servers and clients. We have begun exploring the use of combined blockchain/Linked Data architectures within our own LinkChains work (Third and Domingue 2017).

Decentralisation has the potential to address problems such as those described above, but only when implemented in accordance with sound principles of decentralised data management. The problem we see for the research community is that there is currently no way of comparing decentralised data solutions as they vary greatly and, as far as we know, no comparison framework or standard exists. Our aim in this paper is to address this issue by presenting our FAIR TRADE framework for assessing decentralised data solutions, based on relevant dimensions for decentralization, and clear principles for best practice. Our framework builds upon the FAIR data principles (Wilkinson et al. 2016) which outline how scientific data should be managed, and adds dimensions related to decentralised data. Note that, just as the FAIR principles are focused on the *management* of data, so too are the TRADE principles we outline. More detailed dimensions of evaluation relating to *use* of data systems, such as performance, are not considered here.

The rest of this paper is organised as follows. In section 2 Background and Related Work, we first discuss the meanings of key terms and describe the FAIR data principles, and follow by describing a number of decentralised and distributed approaches to data management including blockchains, Solid and Linked Data Fragments, and our own decentralised approach to data management, LinkChains. Section 3 then describes our FAIR TRADE framework for assessing decentralised data approaches, before in section 5 testing our framework through the evaluation of six different data platforms. In the final section, we summarise and conclude our contributions.

# 2 Background and Related Work

## 2.1 Decentralisation and Distribution

There are a number of different ways in which data can be "spread out" across multiple locations. Each location could contain full copies of the same dataset, or different locations may contain different data. Independently, data at each location may be controlled or coordinated by a single central source, or different locations may be independent. In this paper, we refer to the former as the *distribution* of data, and to the latter as *decentralization* of data. The general theme in the literature appears to be loosely that distribution refers to location and decentralization to independence, but there is no clear consensus on definitions – e.g., Asano et al. (Asano et al. 2018) define decentralization to be when "data are maintained in different sites with autonomous storage and computation capabilities", which would make decentralised data systems a subset of distributed ones as defined by, e.g., Öszu & Valduriez (Özsu and Valduriez 2011) ("a number of autonomous processing elements […] interconnected by a computer network and that cooperate in performing their assigned tasks"). Interpreting these terms independently in the way we do permits greater expressive power, giving the option to describe all four combinations of decentralised or not vs. distributed or not.

Motivations for data distribution include redundancy, to prolong the lifespan of data and preserve it from accidental loss or malevolent attacks, as well as potentially spreading the load of data access beyond a single point of failure. Distribution of data need not be limited to simply storage and copying, but could, for example, also include querying: queries can be computationally expensive, and a solution including distributed query evaluation has the potential for better performance or sustainability than with non-distributed querying. Motivations for decentralised data systems are varied, whether a political wish to increase user control, or a practical need to use independent datasets spread across a network (e.g., the federated Linked Data model).

## 2.2 Autonomy and Identity

When the control of data is not restricted to those who have control over those locations in which it is stored (and, indeed, restricted *from* them), and is instead kept in the hands of particular individuals, e.g., the data owner, a data solution can be said to support *autonomy* of data. This is independent of decentralization and distribution; a fully centralized and non-distributed data system may nonetheless support autonomy if, for example, all data and associated metadata were to be encrypted under the data owners' control. There are a number of aspects of control which could be considered under data autonomy. Control over read/append/modify actions means that an individual can decide how much, if any, of their data is exposed, shared, added to, or changed, for how long and for what purpose. Particularly with personal data, control over usage and analysis may limit some of the recent abuses of data which we highlighted above. The format, metadata, and distribution of a dataset can also be controlled. Autonomy in this sense is an individual-focused concept, relative to particular users. As such, *identity* of users is important to analyzing it.

## 2.3 Trust

Trust in general, of course, is a broad term and includes aspects which are very difficult or impossible to address in technological systems. For example, a system cannot guarantee that data (e.g., about educational qualifications) was not fabricated prior to publication. Other aspects of trust are also covered by other principles or concepts. For example, trust that data is solely under the appropriate individual or institutional control falls under autonomy; integrity in the sense of conformance to a given schema falls under the Reusable principle of the FAIR standard, and so on. But there are some specific aspects of trust independent of those principles, and which specifically relate to decentralised data – in particular, properties relating to provenance metadata and integrity in the sense of content persistence over time. Specifically, what degree of assurance can a system provide that the publisher of a piece of data and the context of publication (e.g., date and time) can be correctly identified, and that claims based on the data are accurately based in the data originally published? In the centralised case, these are often guaranteed by the central data store, and trust in these aspects reduces to trust in the central store or its owners. Without a regulatable central authority to appeal to, these are important aspects to consider independently. (Schneider and Trustworthiness 1999) provides a high-level overview of trust in networked environments.

## 2.4 FAIR Data Principles

The FAIR principles emerged from a workshop held in Leiden in 2014 in recognition of the fact that the scientific community had not paid enough attention to the way digital objects including data are managed (Wilkinson et al. 2016). In particular, a need was recognised to improve the infrastructure associated with the use and reuse of scholarly data. A community of researchers, librarians, publishers and funders setup a group FORCE11 (Future of Research Communications and e-Scholarship) which agreed a set of minimal principles for the stewardship of digital scholarly artifacts. The resulting four FAIR principles are:

- ***Findable*** - data given globally unique persistent identifiers and are described via rich metadata. Metadata also has unique persistent identifiers.
- ***Accessible*** - identifiers can be used to retrieve data and metadata via a standard communications protocol which includes authentication and authorization when necessary.

- ***Interoperable*** **-** data and metadata are represented in a formal, accessible and applicable knowledge representation language. Vocabularies are also used to enable interoperability for both data and metadata. Qualified references are contained in (meta)data to point to other (meta)data.
- ***Reusable*** - data and metadata are described with relevant attributes, clear licenses and detailed provenance information.

The above principles have proved popular with the scientific community and have spawned a number of initiatives such as GO FAIR[1] which links the principles to the European Open Science Cloud. The FORCE11 community now has over 2,600 active members.[2] The FAIR principles are not inherently specific to scientific data management, and, as Mons, et al., (Mons et al. 2017) discuss, have begun to be adopted beyond science.

Wilkinson et. al. (Wilkinson et al. 2017) recently discussed a number of additions to FAIR including privacy protection via containers (based on Linked Data Platform containers)[3] and MetaRecords which return metadata given an HTTP GET request.

# 3 The FAIR TRADE Framework

Evaluation frameworks such as FAIR apply to both centralised and decentralised data solutions with little modification, but, in the decentralised case, we argue that they are insufficient. There are aspects of decentralisation which can have a significant effect on possible use of data, and which should be accounted for in any evaluation framework. To address this, we propose the FAIR TRADE framework for assessing decentralised data solutions. FAIR TRADE is an extension of FAIR: findability, accessibility, interoperability and reusability remain relevant dimensions for evaluation. In addition, we add dimensions of being TRusted, Autonomous, and Distributed and dEcentralised to form TRADE. Assessing a data solution according to FAIR TRADE provides a clear characterisation of its properties in terms of decentralisation and reusability.

While there is no essential need to modify the existing FAIR principles to apply to decentralised data, there are particular aspects worth noting that may require attention in interpreting them. For example, appropriate adaptations to ensure global uniqueness and persistence of identifiers may need to be made in a decentralised scenario to support Findability. The authorisation and authentication are harder in decentralized autonomous contexts, so they need more attention. The principles of Interoperability and Reusability are more straightforwardly agnostic of (de)centralisation status.

**Our contribution**

Decentralisation brings with it a number of further dimensions along which data management can vary, and which relate to the quality of a data source. We propose that the dimensions of trust, autonomy, distribution and decentralisation are effective aspects of data management to consider in decentralised contexts, and describing where in a FAIR TRADE space a data source lies is a useful characterisation of its nature, with adoption of *all* of the principles an indicator of best-practice in decentralised data stewardship.

1. ***TRusted*** – *T1: Data publication metadata can be automatically verified and validated* (e.g., that the publishing individual or organization, and timestamp of publication can be checked).

*T2*: *Claims made on the basis of data contents can be automatically verified* (e.g., that data describes the award of an educational certificate to the data owner). In interpreting this principle in specific contexts, the concepts and mechanisms of Verifiable Claims (as per the W3C Working Group) are useful: trustable access to properties of data, which is privacy-respecting, cryptographically secure and automatically verifiable.

[1] https://www.go-fair.org

[2] https://www.force11.org/community/members-directory

[3] https://www.w3.org/TR/ldp/#ldpc-container

Both of these principles are aided by support for a robust notion of identity, with attestations. This is important for data attribution and claim interpretation, and the various Verifiable Claim roles of holder, issuer, inspector-verifier and identity registry can all be relevant.

2. **Autonomous** – AU1: The owner(s), or authorized controller(s), of a piece of data have control over access to, and use of, data.

AU2: Data relating to personal identity follows AU1 and is trusted in the sense of T2.

Note that autonomy of data is not the same as accessibility under the FAIR principles. One may have access to, and have the credentials to have access to, one's data without having autonomous control over it. For example, personal data stored on the Facebook platform can be accessed by authorised users, but those users, including the person whose data it is, are not in control of it. In general, the principle of Autonomy relates to the legal and moral rights of data subjects over their own data.

3. **Distributed** – D1: In a given network, data is physically stored across some proportion or selection of nodes in that network.

D2: In a given network, the evaluation of data processing tasks can be executed across some specifiable proportion or selection of nodes.

Distribution is not necessarily a simple yes/no question: there are different degrees to which data, and data processing, can be distributed. Analysis of a system according to this principle supports, as appropriate, the characterization of quantified degrees of distribution for both D1 and D2 instead of a binary choice.

4. **dEcentralised** – E1: In a given network, no single node or small set of nodes controls which subset of data contents are held by any node or small set of nodes other than itself.

E2: In a given network, no single node or small set of nodes decides which new nodes can join.

A range of scenarios implementing the Trusted, Distributed and Decentralised principles are outlined in (Third and Domingue 2017), which we have since developed further into the LinkChains approach described in section 5.

# 4 Examples of Decentralisation

## 4.1 Distributed Ledges and Distributed Data Storage

It is important to distinguish between the terms 'distributed ledgers' and 'blockchains', which are often incorrectly used as synonyms. Distributed ledgers are replicated, shared and synchronised digital data geographically dispersed over multiple sites possibly including multiple institutions. A peer-to-peer network is required for communication and a consensus algorithm ensures replication and synchronisation across the multiple nodes. A key difference between applications that run on standard platforms and those that run on top of distributed ledgers is the way that data is stored and managed. Rather than connecting from a device (e.g. a mobile phone) to a central server, which holds all the required data (including private data), every player or volunteer in the network participates in ensuring that the whole network contains multiple copies of all the data. This changes a fundamental dynamic. The notion of centralised control disappears completely, rather data and computation are evenly owned, controlled and shared across the peer network. A blockchain is a specific type of distributed ledger where an ever-growing list of records, called blocks, are linked together to form a chain – hence the term 'blockchain'. The first blockchain was conceived by Satoshi Nakamoto in his white paper (Nakamoto 2008) as the basis for Bitcoin, the most famous blockchain based crypto-currency. The main idea behind Bitcoin was to create a currency specifically for the Internet rather than (as is the case in all fiat currencies) mapping an originally physical currency to the global communications infrastructure. Research produced by the University of Cambridge estimates that in 2017, there were 2.9 to 5.8 million unique users using a cryptocurrency wallet, most of them using Bitcoin.[4] The first issue that arises with Internet-based currencies is what is called the 'double spending problem'. This is the case when a digital 'coin' is spent, by an individual, for some service or good, and then the same coin is spent again by the same individual, for example, by copying or duplicating

[4] https://www.jbs.cam.ac.uk/faculty-research/centres/alternative- finance/publications/global-cryptocurrency

the relevant data. Blockchains address this problem by providing an immutable public ledger of all historical transactions. Once processed and stored within a block, and sufficient subsequent blocks have been confirmed, a transaction cannot be altered even by the transaction owners. Figure 1 shows a blockchain containing three blocks. Starting from the right, the newest block, each block points to its predecessor using a hash function. Transactions are stored in a Merkle Tree (Merkle 1980) - a tree of hashes - where the leaf nodes contain the transactions. This structure allows for efficient retrieval and ensures the veracity of the individual transactions in addition to the block – if a transaction is altered, then the hash will no longer be valid.

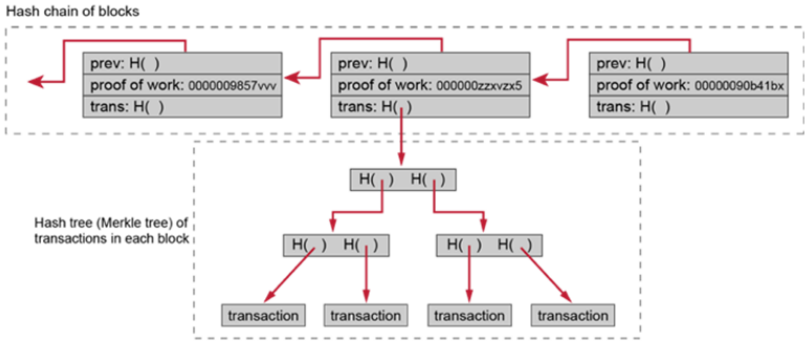


**Figure 1: The hash links, proof-of-work nonce and Merkle tree of transaction data within a blockchain**

The proof of work consensus mechanism which involves solving a cryptographic puzzle before anyone else has led to the growth of the computing power and electrical consumption associated with blockchain networks. Estimates are that by 2020 the Bitcoin network will expend as much electricity as Denmark.[5]b This has led to several blockchain platforms exploring other consensus mechanisms such as:

- ***Proof of stake*** – where the chances of being selected to produce the next block depend on the value of a 'stake' stored by a miner in a specific location. Variants of this take into account the 'age' of the stake.
- ***Proof of capacity*** – rather than the chances of being selected being related to the amount of computing power, as for proof of work, here the probability is related to the amount of storage a miner holds.
- ***Proof of burn*** – sending coins to an irretrievable address ('burn') gives one the right to be selected. The chances of being selected to mine the next block are related to the value of the burn.
- ***Proof of elapsed time*** – Intel has produced a special processor capability to implement a mechanism which relates elapsed time to the probability of being selected.

After Bitcoin, Ethereum (Buterin 2013) is the best known blockchain platform. Rather than serving as a platform for a cryptocurrency the underlying aim for Ethereum is to be an open blockchain platform to support the development and use of decentralised applications. Unlike Bitcoin, the programming language available on the Ethereum platform is Turing complete, so that general applications can be run on what the founders call a 'world computer'.

At the core of the Ethereum concept are two types of accounts:

- ***Externally Owned Accounts (EOAs)*** which are controlled by private keys. A private key is a cryptographic mechanism allowing for individuals to sign a transaction which has been secured by a corresponding public key. EOAs are controlled by individual users or organisations.
- ***Contract Accounts,*** also termed 'Smart Contracts' are controlled by contract code and are activated by EOAs.

[5] https://motherboard.vice.com/en_us/article/aek3za/bitcoin-could-consume-as-much-electricity-as-denmark-by-2020

When *ether*, the currency used within Ethereum, is sent from an EOA to a Contract Account, the contained program is executed. This can result in further transactions and payments and additional Smart Contracts being invoked. Through these chains of invocation, connected Smart Contracts form the basis of Ethereum applications called 'dApps' (distributed applications). A number of high-level languages exist for Smart Contracts including Solidity (similar to C/JavaScript), Serpent (similar to Python) and LLL (a low-level Lisp-like language).

One Smart Contract that has attracted attention recently is ERC721 ("Ethereum Request for Comments" #721).[6] This Smart Contract was developed after the Crypto Kitties[7] game became very popular. Players of Crypto Kitties can collect, trade and breed cats, with all activity running on the Ethereum blockchain. At its height, the platform raised $12M of venture capital and the most popular cats were valued at over $100,000. ERC721 extends ERC 20, a Smart Contract for representing tradeable tokens (such as coins). The key contribution of ERC721 is to represent *non-fungible* tokens - tokens which are each unique and therefore not interchangeable. Below we outline how we have extended this Smart Contract in our own work.

The Interplanetary File System (IPFS)[8] is a peer-to-peer distributed file system which is in some ways analogous to the Web, but which uses BitTorrent techniques for exchanging data. Each stored file is indexed by its hash with the indexes stored on network nodes. Human readable decentralised file naming is supported through IPNS analogous to DNS. The benefits brought by IPFS include content-based addressing, increased speed of data delivery over networks since large files can be transported in parallel, easy replication of valuable data, and network resilience in areas of low connectivity. IPFS has often been used together with blockchains as a data storage area and one recent joint venture is Filecoin[9] whereby peer data hosters are paid.

## 4.2 Verified Claims and Self-sovereign Identity

The Verifiable Claims Working Group (VCWG) aims to make the verifiable exchange of claims easier and more secure on the Web – in particular, how to make claims cryptographically secure, privacy respecting, and automatically verifiable. The main roles in the use of claims and their relationship are outlined in Figure 2, taken from (Sporny and Longley 2017)):

- ***Holder*** - the owner or controller of a number of verifiable claims, for example a student with qualifications or a citizen with personal attributes (e.g. age) or rights (e.g. to work).
- ***Issuer*** - creates verifiable claims, each connected to a specific subject, and then transmitted to claim holders. Examples here include universities (for qualifications) and governments.
- ***Inspector*** verifier - processes all received verifiable claims. Examples include employers and staff managing national borders.

***Identifier registry*** - mediates the creation and verification of entities about which claims are made. Examples may include university records, organisational employee databases and governmental citizenship data stores. It is acknowledged that distributed ledgers can play the role of identifier registries in decentralised contexts. Standard identity management systems are based on centralised authorities such as certificate authorities or domain name registries. These are in some cases perceived as contributing to some of the problems outlined above, that is, contributing to the over centralisation of data. A partial community response to this has centred around the notion of 'self-sovereign identity' whereby users own, control, and manage their data. The main mechanism for this is through the following two main steps. Firstly, using a special app - an identity wallet - the user generates a public/private key pair which is unique and stored locally. Secondly, users request attestations using their public key from authorities and these are digitally

[6] http://erc721.org/

[7] https://www.cryptokitties.co/

[8] https://ipfs.io/

[9] https://filecoin.io/

signed by and stored on the wallet as well. An identity check from an identity authority may be conducted using mobile phone and passport to verify a user's identity. A blockchain can serve as a public trusted store for public keys. The main benefits associated with self-sovereign identity are that users maintain ownership and control over their data. Users can share attestations which can be shaped to minimise the exposure of private data. For example, a citizen may have an attestation, from an appropriate governmental agency, that they are 'over the legal drinking age' independent of any stored record of birth date.

Building on distributed ledger technologies, such as blockchains, the W3C Decentralised Identifiers Community Group have recently generated a data model and syntax for Decentralised Identifiers (DIDs) (Reed et al. 2018). DIDs are a new type of identifier for verifiable self-sovereign digital identity under the control of a DID subject, independent from any centralized authority. Each DID Document is comprised of three parts. *Cryptographic material* enables documents to be authenticated through a suite of *authentication suites. Service endpoints* support the provision of decentralised identity management services using the data contained in the DID.

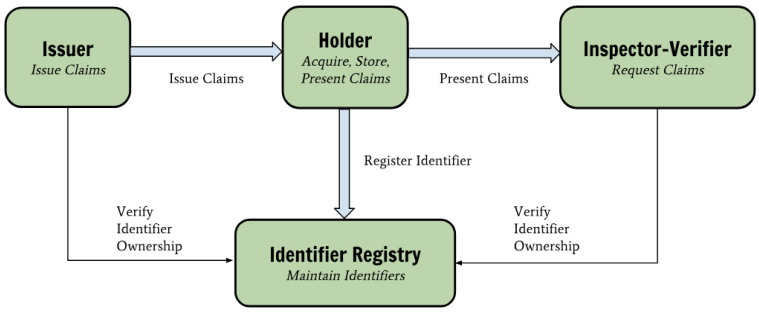


**Figure 2: An example type of scenario for the use of verifiable claims. Taken from** (Sporny and Longley 2017)

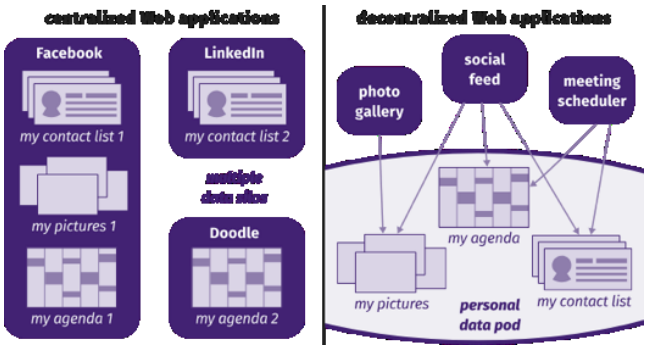


**Figure 3: An overview of Solid concepts (taken from (Verborgh 2017))**

## 4.3 Decentralised Linked Data Approaches

The Linked Data model supports, by design, the capacity for distributed or decentralised data systems. A single SPARQL query may, through the SERVICE keyword, pull data from multiple independent sources and aggregate it to produce a single set of results. Linked Data approaches in general, however, are agnostic with regard to distribution and decentralisation and can be used to implement many variations along these spectra. In terms of approaches specifically designed with decentralization in mind, we discuss Solid and Linked Data Fragments.

Solid (Lee 2018b) is a new initiative from the inventor of the Web which aims to enable the decentralisation of personal data. As can be seen in Figure 3, each Solid user has their own personal data which is called a data pod. Users are free to choose who hosts the data - whether to host personally or with

a third party of choice. In the Solid scenario, applications are decoupled from the data they consume, interacting through APIs, with user permission, avoiding vendor lock-in and allowing users to easily switch between applications and servers without loss of service. An additional benefit is that newcomers can innovate and develop new applications for users to try out. As stressed in an open letter (Lee 2018c), Solid is driven by the principle of "personal empowerment through data".

According to Verborgh et al., (Verborgh et al. 2016) a Linked Data Fragment is a subset of an RDF based knowledge graph computed by some means. As stated by the authors the problem with most approaches to obtaining Linked Data Fragments outside of a centralized context is that the burden of computation rests entirely with the client, for example, if the data is downloaded, or with the server, for example, using SPARQL, as shown at the top of Figure 4. As shown at the bottom of Figure 4, Triple Pattern Fragments (Verborgh et al. 2016) allow the computation to be shared between a server and client. In essence the approach lowers the burden on RDF servers since the only computation required is on the simple matching of triples. The Triple Fragment Server returns matching triples in reasonably sized fragments (100 triples approximately) and, in addition, metadata and controls are returned. The metadata includes an estimate of the number of triples that match the given triple pattern. Controls enable clients to retrieve further triples from the same knowledge graph. A Linked Data Fragments client supports standard SPARQL queries over any returned triples. Triple Pattern Fragments make it easy to query across multiple data stores at once - much as with SPARQL's SERVICE keyword, but with the bulk of the computational work of aggregation moved from server to client. Triple Pattern Fragments offer a specific advantage for decentralised querying compared to SPARQL, in that it is no longer *required* to know in advance which pieces of data are to be found on which federated endpoint.

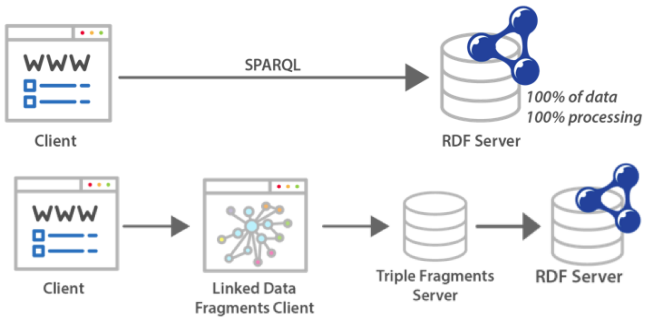


**Figure 4: At the top, a standard SPARQL scenario where an RDF server is responsible for hosting and processing all the data. At the bottom, a combination of a Linked Data Fragments client and a Triple Fragments server allows the burden of computation to be shared between a client and server.**

## 4.4 LinkChains

In designing LinkChains, our overall goal has been to bring together the benefits accrued from LD approaches together with decentralised approaches, as embodied by distributed ledgers and IPFS. As can be seen in Figure 5, the LinkChain architecture has three main parts: a user-controlled private storage area; a decentralised public and private data storage area; and a blockchain.

We currently use Solid personal data pods to store private data solely under user control (complete control over the pod-stored data), although LinkChains will talk to any Linked Data Platform.[10] Read/write access to Solid pods is granted through the Solid auth client (Lee 2018b). Non-private or encrypted data is stored within a decentralised data storage system, as opposed to directly on a blockchain, as most public blockchains require payment for data storage and do not store large volumes of data efficiently. We use a blockchain instead to record and verify claimed data attributes, and to facilitate the transfer of data as an asset.

[10] https://www.w3.org/TR/ldp/

A LinkChain Smart Contract enables data on the blockchain to be written or queried. The key functionality provided through our Smart Contract is the ability to store LinkChain Trust Tokens (LCTT) - tokens verifying a set of attributes of a Linked Data set. In particular, a LCTT provides signed verification that declared attributes hold for a Linked Data set. LCTT are created by extending the ERC 721 token standard described above. The main parts of this LinkChain architecture are controlled through a Web/mobile application or a client using the APIs and interfaces provided by the LinkChain platform.

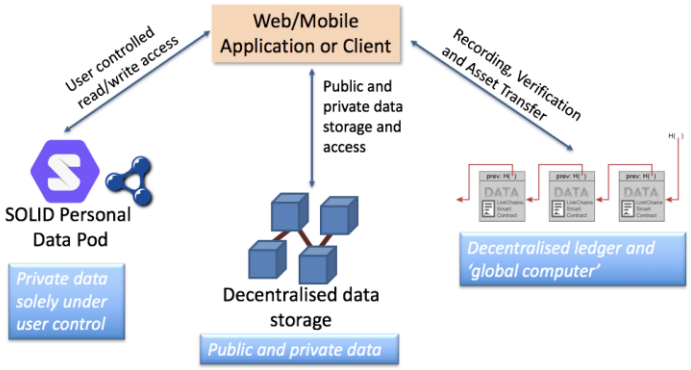


**Figure 5: The overall LinkChain architecture which comprises of three main parts: storage of private data under user control; decentralised public and encrypted private data; and a blockchain for recording, verification and asset transfer.**

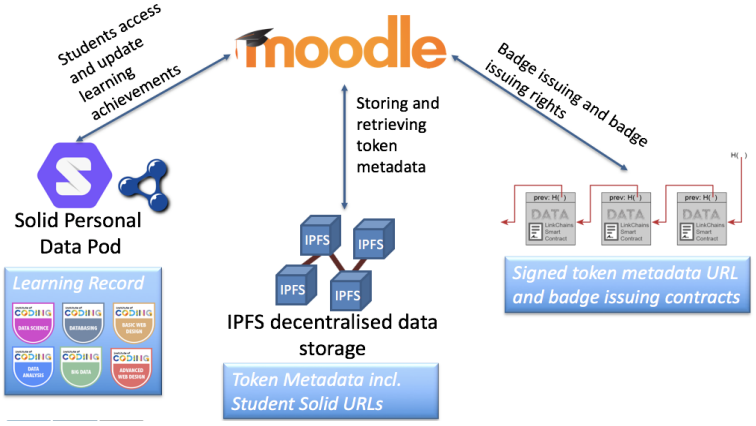


**Figure 6: An instantiated LinkChain architecture supporting the storage of personal learning records including Open Badges.**

Figure 6 shows how we instantiated the above generic LinkChain architecture to store and verify lifelong learning records. This ongoing work is a part the Institute of Coding[11], an initiative from the UK government to tackle the UK digital skills crisis.[12] When a student completes a particular assignment using the Open University's Virtual Learning Environment (VLE), an Open Badge (based on the IMS Open Badge standard)[13] is automatically issued which students can consequently claim and have stored within their Solid data pod. Additionally, a digitally signed hash of the badge is encoded within an LCTT placed

[11] https://instituteofcoding.org/

[11] https://londonlovesbusiness.com/institute-of-coding-launches-to-tackle-uk-digital-skills-crisis/

[13] https://www.imsglobal.org/cc/statuschart/openbadges

onto the LinkChain blockchain and a token metadata file is stored on IPFS. Placing the metadata on IPFS rather than onto the blockchain reduces cost and increases the overall data handling efficiency. Figure 7 shows a screen snapshot of the student interface used when the system was deployed in our 2018 Summer of Code online school.[14] We can see in Figure 7 that the student has selected her Super Badge, given for successfully completing all nine assignments, and can view the details of the blockchain badge representation. Figure 7 also shows that each element of the Open Badge, including *assertion*, *recipient* and *evidence*, has its own distinct blockchain representation encoding data such as the date of issue and whether the badge has been revoked or has expired.

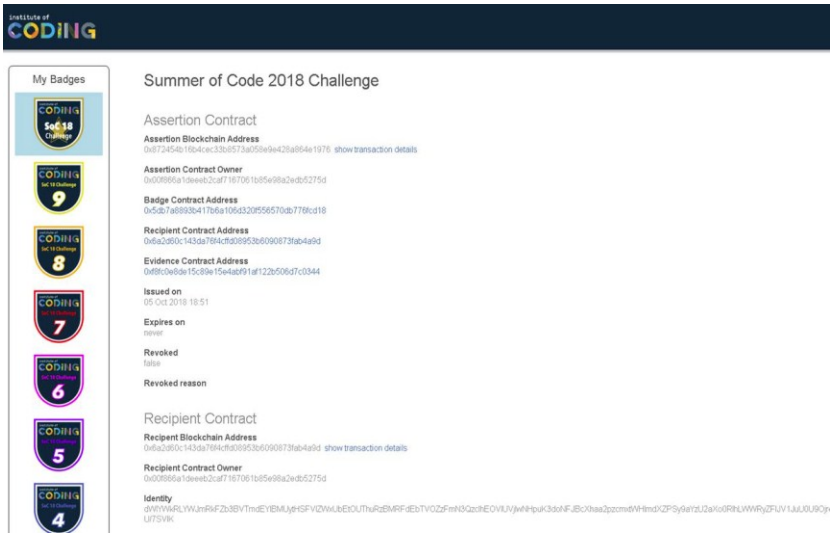


**Figure 7: A portion of the interface for OU students viewing their blockchain badges on the 2018 Summer of Code course.**

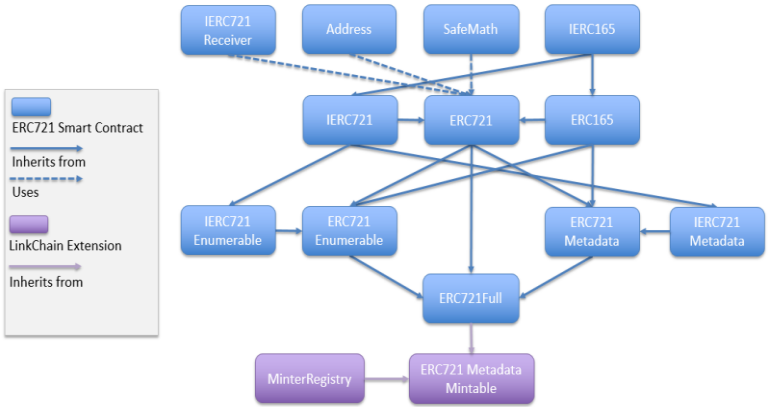


**Figure 8: Our extensions to the ERC 721 Smart Contract to support our LinkChains implementation**

Figure 8 shows ERC721, including the extension we created for issuing the non-fungible tokens within our LinkChain platform. Each rectangle in the image represents a distinct Smart Contract and the set of blue rectangles together form the ERC721 Smart Contract Template labelled ERC721Full. This template is part of OpenZeppelin[15], a framework of reusable smart contracts for Ethereum and Ethereum Virtual Machine

[13] http://www.open.ac.uk/about/teaching-and-learning/esteem/projects/themes/supporting-students/summer-code

[14] https://github.com/OpenZeppelin/openzeppelin-solidity/blob/master/contracts/token/ERC721/ERC721.sol

(EVM) based blockchains (i.e. blockchains which use the EVM as an execution platform). ERC-721 defines a minimum interface a smart contract must implement to allow unique tokens to be managed, owned, and traded. IERC721 is an interface which defines the functions used in the smart contracts and is used as a way to implement those functions. The "address" smart contract is used to separate account addresses and smart contract addresses. The SafeMath smart contract makes sure that the mathematical operations performed in a smart contract are safety checked which can revert on error. ERC165 standard interface detection is used to expose the interfaces that an ERC721 smart contract supports. The ERC721 Enumerable extension provides functionalities to sort through the tokens easily, whereas the Metadata extension enables a token to have a contract name, symbol and some extra data that makes it unique.
The two smart contracts MinterRegistry and ERC721MetadataMintable (highlighted in purple) together form the core of our LCCT token implementation. MinterRegistry contains a list of addresses which are permitted to mint, i.e. are allowed to issue new ERC721 tokens. ERC721MetadataMintable issues LCCT tokens upon receiving legitimate minter requests in the form of metadata representing the address of the minter, the address that the tokens created are to be sent to, the address of the token contract and the token ID. On ERC721MetadataMintable being invoked, the smart contract authenticates the minting address and issues the LCCT token to the specified address.

# 5 Evaluation according to FAIR TRADE

We evaluate these principles by considering a number of decentralised data systems and identifying which of the principles, and subprinciples, apply in each case, and to what degree. The goal is to test their *expressive power*: in applying these principles, can we describe significant differences with regard to various systems and approaches? The choice of systems to evaluate is diverse, including systems with significantly different goals and designs, deliberately to illustrate this expressive power.

## 5.1 Solid

Solid accrues many of its positive scores with respect to the FAIR TRADE framework due to the fact it is a set of conventions and tools based on Linked Data principles and more generally as far as possible on existing W3C standards and protocols. Specifically, the use of URIs supports *findability* through the provision of a unique and persistent identifier. *Accessibility* is supported through the return of RDF based metadata, understandable to both humans and machines, available through HTTP requests and SPARQL queries. Metadata in a formal knowledge representation language such as RDF also supports *interoperability*. The PROV-O provenance vocabulary[16] and the Linked Data Rights vocabulary[17] can support *re-usability*.
We take the relevant network, in the case of Solid, to be the collection of all data pods. *Distribution* and *decentralisation* of data are therefore satisfied for all subprinciples. Each is a distinct per-user Web server allowing users in principle to be completely *autonomous* in terms of data control (AU1), but without a stronger model for identity, AU2 does not hold. Solid does not satisfy any aspects of the *trusted* dimension. This highlights the fact that there is no inbuilt mechanism for verifying that any claims made within a data pod are true. The Verifiable Claims data model can be used to express claims, but additional infrastructure would be required to support verification.

## 5.2 DBpedia Infrasructure

DBpedia (Lehmann et al. 2015) is, at its core, a machine-readable formal representation of the various Wikimedia projects including Wikipedia. Data within DBpedia is served as Linked Data, enabling semantic query processing. The English version of the DBpedia knowledge base contains over 4.5 million objects, over 4.2 million of which are classified within the DBpedia ontology. DBpedia is not, in itself, of course, a

[16] https://www.w3.org/TR/prov-o/

[17] http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/

data platform; it is a dataset. Nonetheless, there is an infrastructure surrounding its creation, management, and publication, including a central DBpedia site[18] and a network of mirrors. We discuss this infrastructure here.

As it is founded on Linked Data, DBpedia is positively assessed against all the FAIR criteria. Every DBpedia data item has a unique dereferencable URI which contains data and metadata represented in RDF satisfying the *findability*, *accessibility* and *interoperability* dimensions. All data in DBpedia is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License supporting the *re-usability* principle.

*Trust* is generated from the fact that the data has been extracted from Wikipedia, and that Wikipedia generates trust for the information it holds from its inbuilt consensual and social mechanisms.[19] However, the "automatic" aspects of T1 and T2 are not satisfied. DBpedia data is centralised, and ultimately controlled by the DBpedia Board of Trustees,[20] which currently has eight members; data is not *autonomous* with respect to contributors or users in the senses of AU1 and AU2. DBpedia provides tools to make it easy to maintain a live mirror,[21] however, synchronisation is one-way (central data is pushed to the mirrors); D1 and E2 are satisfied, but D2 and E1 are not.

## 5.3 Linked Data Fragments

As with the systems described above, Linked Data Fragments passes all the FAIR criteria as it is based on Linked Data, but only partially satisfies the new criteria. As with Solid, there are no inbuilt mechanisms to facilitate *trust* of any provenance for (T1), or claims made about (T2), returned data. Following the core aim of Linked Data Fragments, the platform facilitates the distribution of data between clients and servers - primarily to enable the sharing of computation association with processing. Linked Data Fragments certainly supports both D1 and D2 with regard to *distribution*, and E1 and E2 with regard to *decentralisation*. With regard to *autonomy*, data on a Linked Data Fragments server is controlled by whomever controls the server. and there is no standard means of controlling data visibility. AU1 does not, therefore, hold, nor does AU2: there is no consideration of user identity in the standard at all.

## 5.4 Bitcoin

Bitcoin core can be used to find transaction/exchange data. But, the metadata or message embedded in the OP_RETURN part of the bitcoin transaction does not lend itself to findable data. Bitcoin as a platform contains no internal indexing or search mechanisms - all discovery is down to external tools which are fixed, or domain specific, or proprietary (e.g. Blockchain.com)[22]; thus transaction/exchange data is *findable*, but data embedded within a Bitcoin transaction is not. As Bitcoin is a public blockchain, all data is *accessible* from the Bitcoin blockchain at any time. No format is prescribed for representing embedded data on Bitcoin, nor is there support for vocabularies; embedded data in a Bitcoin transaction is in general is not *interoperable*. The fact that every Bitcoin transaction including the embedding of data is signed by a private key means that data has at least some verified provenance. This makes embedded data in a Bitcoin transaction *re-usable* to some extent, although it has no standard license, or licensing infrastructure.

Bitcoin miners are distributed over the world and every node has a copy of the entire blockchain. The combination of peer data replication, the proof of work consensus mechanism, and immutability means that Bitcoin data can be *trusted* – neither provenance metadata (T1) nor data contents (T2) can be modified once published. *Autonomous* control of data directly on-chain is limited; data cannot be deleted from the

[18] http://dbpedia.org

[19] https://en.wikipedia.org/wiki/Wikipedia:Consensus

[20] https://wiki.dbpedia.org/board

[21] https://github.com/dbpedia/dbpedia-live-mirror

[22] https://www.blockchain.com/explorer

blockchain, although some control over visibility can be implemented using encryption. Hashes with pointers to off-chain data offer greater potential for autonomy, as does the use of multiple anonymous accounts via a management wallet, but these, particularly the former, require infrastructure going beyond Bitcoin itself. We cannot therefore say that it satisfies either of the autonomy subprinciples The Bitcoin ledger is globally *distributed*: every node has a copy of the full blockchain (D1) and every node executes every script (D2), and fully *decentralised* (E1 and E2). There have been concerns within the Bitcoin community that an individual or consortium of miners could take over the network by gaining more than 50% of the total hashing power and thereby the ability to rewrite history. Thus far, this has not happened; at the time of writing, all mining groups have less than 20% control of the overall network.[23]

## 5.5 Ethereum

As Ethereum and Bitcoin share similar features, its evaluation results according to FAIR TRADE are similar to Bitcoin. A noteworthy difference is with regard to *autonomy*: the availability of smart contracts gives Ethereum increased power in terms of what the blockchain network itself can support; the autonomy subprinciples themselves may potentially be easier to support on Ethereum with a self-sovereign identity and Verifiable Claims infrastructure using smart contracts. Nonetheless, further work, and off-chain infrastructure, would be required in order to support both AU1 and AU2 fully.

## 5.6 LinkChains

LinkChains is designed to combine the user-focused Linked Data design of Solid with a blockchain-backed trust layer, and thus inherits traits from both Solid and Ethereum. That it meets more of the FAIR TRADE criteria than other systems here is not a surprise; a FAIR TRADE analysis played a role in the development process. Rather, it provides an example of how this framework can focus efforts by making explicit the principles for high-quality decentralised data management. The key element *not* yet present in the LinkChains design for full FAIR TRADE compliance is support for *autonomous* identity, although we are currently exploring solutions for self-sovereign identity approaches to address this gap.

**Table 1: Evaluation according to FAIR TRADE**

| | Solid | DBpedia | LinkChains | Bitcoin | Ethereum | Linked Data Fragments |
|---|---|---|---|---|---|---|
| Findable | Via LD | Yes | Via Solid | No internal index or search | Similar to Bitcoin. | Yes |
| Accessible | Via LD | Yes | Via Solid | Yes | Yes | Yes |
| Interoperable | Via LD | Yes | Via Solid | Data in arbitrary format | Data in arbitrary format | Yes |
| Re-usable | Via LD | Yes | Via Solid | No licensing. | No licensing. | Yes |

[23] https://www.blockchain.com/en/pools

| | | | | | | |
|---|---|---|---|---|---|---|
| Trusted | Neither T1 nor T2. | Neither T1 nor T2; not automatic. | T1 and T2, via blockchain. | T1 and T2, by blockchain design. | T1 and T2, by blockchain design. | Neither T1 nor T2. |
| Autonomous | AU1, not AU2. | Neither AU1 nor AU2. | AU1 via Solid, not AU2. | Neither AU1 nor AU2. | Neither AU1 nor AU2. | Neither AU1 nor AU2. |
| Distributed | D1 and D2. | D1, not D2 | D1 and D2. | D1 and D2 | D1 and D2. | D1 and D2 |
| Decentralised | E1 and E2. | Not E1. E2. | E1 and E2 | E1 and E2. | E1 and E2. | E1 and E2. |

# 6 Conclusions

As we stated at the start of this paper, a growing number of societal concerns have been raised over the last few years related to how the centralisation of data has led to the loss of user control and the dangers associated with data misuse. A number of different technical communities associated with peer-to-peer and Web technologies have been responding to this in part with a variety of new approaches and platforms to decentralising data. The contribution of this paper is to provide a framework to begin to compare systems, based on an explicit set of principles relating to decentralised data management.

Above, we have assessed a range of data solutions for adherence to the FAIR TRADE principles. As well as a measure of quality in data stewardship, the exercise of assessment showed that these principles, even when not fully followed, provide a good characterisation of the properties of the data or platform being assessed. One can see at a glance, for example, that the DBpedia project is concerned with providing a common resource -- concerns relating to autonomy do not apply when the goal is to elicit effectively *donations* of information for public use, but those relating to trust, openness and reuse do apply, with distribution only as required to ensure the accessibility of data. The Bitcoin and Ethereum platforms meet most of the TRADE standards, except that of being Autonomous, but not FAIR, reflecting their development from cryptocurrencies and, in Ethereum's case, distributed computation - their use as decentralised data platforms coming later. Solid and Linked Data Fragments are evaluated to be very similar, reflecting their shared concerns and foundations on Web data standards. It is perhaps unsurprising to see that the platforms which most fit the FAIR TRADE principles are those stemming from a field in which the focus is on data publication and standards. LinkChains, although a work in progress, is being designed to be FAIR TRADE from the outset, and is based on a combination of other technologies, each of which follows some of the principles. In particular, combining Solid's focus on individual data with the possibilities for trusted and verifiable data based on blockchains, distributed identity and Verifiable Claims has the potential to provide a flexible platform with best-practice principles in data management in the decentralised sphere. Note that these systems did not stretch the descriptive capacity of the FAIR TRADE principles: as noted earlier, each of the presented principles can cover a range of possible useful or interesting behaviours which a distributed data solution could display. As novel technologies arise, we expect this framework to continue to provide a useful common vocabulary for assessing and comparing decentralised data systems. By taking a forward-looking approach to data management standards in a decentralised context, we aim to contribute to the establishment of best practice and well-motivated technical approaches early in the development of the field, at a time when their adoption is significantly easier than in a more entrenched technical landscape.

## REFERENCES


Asano, Y., Hidaka, S., Hu, Z., Ishihara, Y., Kato, H., Ko, H.-S., Nakano, K., Onizuka, M., Sasaki, Y., and Shimizu, T., 2018. A View-based Programmable Architecture for Controlling and Integrating Decentralized Data. *arXiv preprint arXiv:1803.06674*.

Bureau, C. F. P., 2015. *Data Point: Credit Invisibles* [online]. Available from: https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf.

Buterin, V., 2013. *Ethereum White Paper* [online]. Available from: https://github.com/ethereum/wiki/wiki/White-Paper [Accessed 10 Dec 2018].

Curran, D., 2018. Are you ready? Here is all the data Facebook and Google have on you. *The Guardian* [online], 2018. Available from: https://www.theguardian.com/commentisfree/2018/mar/28/all-the-data-facebook-google-has-on-you-privacy [Accessed 10 Dec 2018].

Fact, F., 2018. Windrush generation: what's the situation? [online]. Available from: https://fullfact.org/immigration/windrush-generation/ [Accessed 10 Dec 2018].

Grierson, J., 2018. Windrush row: 63 people could have been wrongly removed. *The Guardian* [online], 2018. Available from: https://www.theguardian.com/uk-news/2018/may/15/windrush-row-63-people-could-have-been-wrongly-removed-says-javid [Accessed 10 Dec 2018].

Guardian, T., 2015. The Cambridge Analytica Files. *The Guardian* [online], 2015. Available from: https://www.theguardian.com/news/series/cambridge-analytica-files [Accessed 10 Dec 2018].

Lee, T. B., 2018a. Interview in Vanity Fair. [online]. Available from: https://www.vanityfair.com/news/2018/07/the-man-who-created-the-world-wide-web-has-some-regrets [Accessed 10 Dec 2018].

Lee, T. B., 2018b. SOLID Project Website. [online]. Available from: https://solid.mit.edu/ [Accessed 10 Dec 2018].

Lee, T. B., 2018c. One Small Step for the Web. [online]. Available from: https://www.inrupt.com/blog/one-small-step-for-the-web.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., and Auer, S., 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6 (2), 167–195.

Merkle, R. C., 1980. Protocols for public key cryptosystems. *In*: *Security and Privacy, 1980 IEEE Symposium on*. IEEE, 122.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., and Wilkinson, M. D., 2017. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37 (1), 49–56.

Nakamoto, S., 2008. Bitcoin: A peer-to-peer electronic cash system.

Özsu, M. T. and Valduriez, P., 2011. *Principles of distributed database systems*. Springer Science & Business Media.

Reed, D., Sprony, M., Longley, D., Allen, C., Grant, R., and Sabadello, M., 2018. Decentralized Identifiers (DIDs) v0.11 Data Model and Syntaxes for Decentralized Identifiers (DIDs). *W3C* [online]. Available from: https://w3c-ccg.github.io/did-spec/ [Accessed 10 Dec 2018].

Schneider, F. B. and Trustworthiness, C. on I. S., 1999. *Trust in cyberspace*. National Academy Press Washington, DC.

Sporny, M. and Longley, D., 2017. *Verifiable Claims Data Model and Representations* [online]. Available from: https://www.w3.org/TR/2017/WD-verifiable-claims-data-model-20170803.

Third, A. and Domingue, J., 2017. LinkChains: Exploring the space of decentralised trustworthy Linked Data.

Verborgh, R., 2017. Paradigm shifts for the decentralized Web. [online]. Available from: https://ruben.verborgh.org/blog/2017/12/20/paradigm-shifts-for-the-decentralized-web/ [Accessed 10 Dec 2018].

Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., and Colpaert, P., 2016. Triple Pattern Fragments: a low-cost knowledge graph interface for the Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37, 184–206.

Walport, M., 2015. *Distributed Ledger Technology: beyond block chain* [online]. Government Office for Science. Available from: https://pubs.acs.org/doi/pdf/10.1021/acsaem.8b00240%0Ahttp://pubs.acs.org/doi/10.1021/acsaem.8b00240.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., and Bourne, P. E., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3.

Wilkinson, M. D., Verborgh, R., da Silva Santos, L. O. B., Clark, T., Swertz, M. A., Kelpin, F. D. L., Gray, A. J. G., Schultes, E. A., van Mulligen, E. M., and Ciccarese, P., 2017. Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Computer Science*, 3, e110.