technical report

Applying blockchain technology in global data infrastructure

ODI-TR-2016-001

Open Data Institute

Table of contents

Authors: James Smith, Jeni Tennison, Peter Wells, Jamie Fawcett, Stuart Harrison

Editor: Anna Scott

Design: Christie Brewster

Executive summary

Blockchains, or 'distributed ledgers', are part of a new area of technology that is generating a lot of interest. They have the potential to become an important component of our global data infrastructure.

In this report, we present an overview of Blockchain technology and issues that come with it, for a non-technical audience who seek to understand the potential of distributed ledgers and blockchains in a commercial or policy context. We focus on non-financial use cases, both to avoid duplication of other work and to explore the wider impact of the technology.

Some of the key areas covered in this report are:

- the basics of blockchain and distributed ledger technology
- the landscape of use cases and applications
- how blockchains will need to link into our global data infrastructure
- how blockchains will need to grow and scale over time
- the potential privacy implications for personal data in distributed ledgers, and the risk of adding personal information to blockchain systems without careful design
- a practical exploration of building a blockchain system for data storage

We conclude that distributed ledgers are a potentially important area of technology, but that we must avoid being swept up by 'blockchain hype', and remember to focus on solid user needs first, before we choose the technology that we will use.

We recommend further that this is best done by organisations convening across sectors to identify common data infrastructure needs.

Introduction

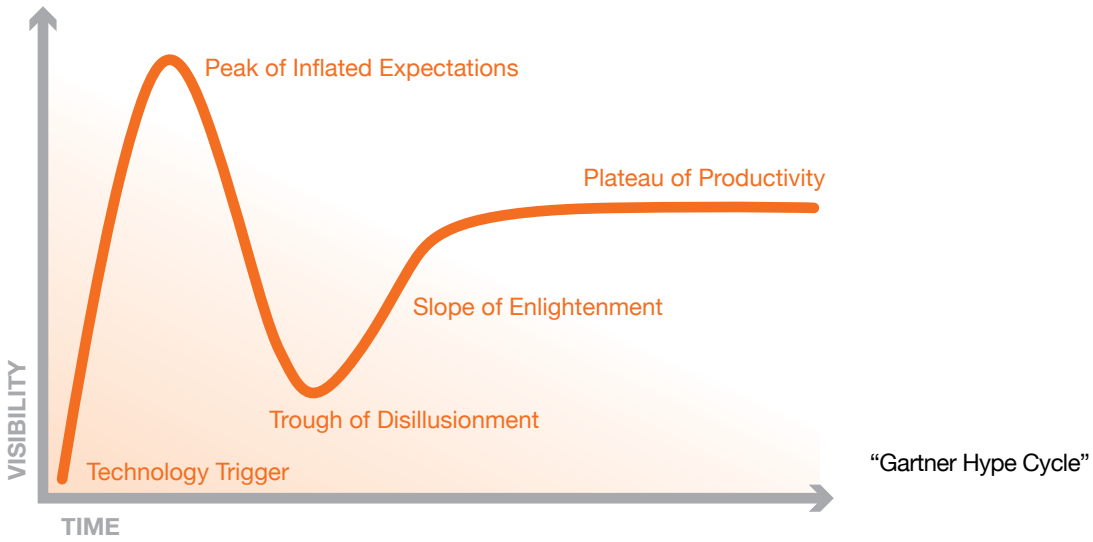
Data is essential for the modern age; it is infrastructure for the whole economy. It underpins public service transformation, business innovation and democratic engagement. It connects sectors.

Data infrastructure includes technology, processes and organisations. The Open Data Institute (ODI) is in the process of developing its design principles for strengthening the data infrastructure we rely on to build tools, products and services that benefit everyone. Sometimes we talk about our data infrastructure like we do our road infrastructure: roads help you get from A to B – data helps you get to a decision. Sometimes we need a super highway or a motorway; at other times a country lane. Data infrastructure needs to be as reliable as necessary and as open as possible.

A new class of data infrastructure technologies has recently emerged, known as 'distributed ledgers'. Blockchains are one specific type of technology in that domain, and though far from the only type, they are receiving a lot of attention. Blockchain technology emerged from the digital currency Bitcoin, and has been hailed as a revolutionary step forward for data storage and the decentralisation of computer systems.

With the support of Deutsche Bank, the ODI has been exploring emerging data infrastructure technologies – including distributed ledgers and blockchains – their potential and the issues they might bring. We have explored the technology landscape, raised questions based on our recent work on data infrastructure and personal data, and begun to investigate technical implementation. Each of these activities has led to new insights into the current state, and potential, of these technologies.

While there are many technologies relevant to data infrastructure, we have focused recently on blockchains because of the hype surrounding them, and a concern that it may lead to negative outcomes if they are used in the wrong ways.



"Gartner Hype Cycle"

Many technologies go through a similar hype cycle; the "peak of inflated expectations" can lead to the "trough of disillusionment", before eventually we reach the "plateau of productivity" (Gartner, 1995). Blockchains and distributed ledgers are definitely on the upwards slope towards the peak, and in our work we hope to help accelerate through the cycle, and rather than increasing the height of the peak, to instead raise the floor of the trough. Good ideas die in the trough, and we want to help those good ideas survive through the cycle to success.

It is only together that we will be able to answer the question of whether blockchains are "as fundamental for forward progress in society as the Magna Carta or the Rosetta Stone" (Swan, 2015, p. viii), whether they are irrevocably bound to the "failed Bitcoin experiment" (Hearn, 2016), or whether the truth lies somewhere in between.

In this report, we hope to help the reader form their own opinion on that emerging question, and provide a roadmap for further exploration.

The foundations of blockchain technology

Blockchains provide a way to store information so that many people can see it, keep a copy of it, and add to it. Once added, it is very difficult to remove information. This can reinforce trust in a blockchain's content.

At its heart, a blockchain is simply a new kind of database, with a unique set of properties. Rather than being stored in a single location, this database is shared across the Internet, with many people holding a copy of it. The synchronisation between all these different copies – so that there is still a single recognised truth – is one of the unique aspects of blockchain technology.

In his article 'Avoiding the pointless blockchain project,' Greenspan (2015) outlines these properties as follows:

- **Shared read:** blockchains are a structured data store that many people can read
- **Shared write:** as well as read, many people can write data into the database
- **Absence of trust:** the different writers do not have to trust each other not to manipulate the shared database state (i.e. you cannot change information I have added)
- **Disintermediation:** there is no need for a trusted intermediary to enforce access control (or none is available)
- **Transaction interaction:** records in the database depend on and link to each other
- **Validation rules:** the rules around database transactions are well-defined, such that anyone with a copy of the database can validate that it has been maintained correctly

If an application does not need to meet all the above criteria, then it does not require a blockchain solution; there are much more mature technologies available that meet many of the above criteria (including traditional databases) and can also support distributed service and organisational models.

How data is stored in a blockchain

The fact that writers do not need to trust one another not to change information in a blockchain is down to how data is stored in the system. A blockchain stores a series of transactions – which can be data of any sort – in blocks, which get added one after the other in a continuous chain. Importantly, blockchains are considered append-only, or immutable, data stores. Data can be added to the chain, but because each item depends on the one before it, data cannot be removed or changed without recalculating every subsequent transaction.

Because of this, once data is embedded in a blockchain it cannot be altered without that change being detected and potentially rejected by the other nodes in the network. This is useful for data that users need to trust, because it provides a guarantee that the data has not been changed since it was put in the blockchain, but also has other implications, which we discuss below.

Blockchain distribution and data consensus

The disintermediation property arises because blockchains are distributed; there is no central storage location, no 'primary' copy. Blockchains are maintained by a peer-network of nodes; every node has a copy of the blockchain and has equal authority to add to it. Every node publishes that data for other nodes to pick up and use.

The size of the network is important, and ensures the immutability of the database. A single node could change historical transactions and recreate a valid blockchain, but because other nodes in the network check the data before they accept it, it becomes impossible for a single node to make a historical change. More than 50% of the network needs to agree on 'the truth' for it to be accepted. Conversely, there is a danger that if more than 50% of the nodes (or 'miners', in Bitcoin parlance) decide to change history, they can.

As new blocks may be added on the end of the chain by many nodes at once, blockchains need a resolution mechanism to decide which block is accepted by the network. Bitcoin uses a 'proof of work' algorithm, where nodes must prove they have solved a complex cryptographic puzzle, purely as a way of deciding which node gets to add its data first. Other systems use different methods such as the 'proof of stake' algorithm, which requires that nodes prove they own a certain amount of an asset, such as the cryptocurrency of the system. Choosing a resolution mechanism for a blockchain can greatly affect the scalability of the system and how much energy it uses, particularly as 'proof of work' algorithms require a lot of computing power.

‘Distributed ledgers’, ‘blockchains’, or ‘the blockchain’?

Confusing terminology is often used around blockchains, which are themselves a type of a more general ‘distributed ledger technology’ (DLT).

Some people have a tendency to talk about ‘the blockchain’, which normally implies the blockchain that underpins Bitcoin, as opposed to ‘a blockchain’.

The ODI uses blockchain terminology as follows:

- **Distributed ledger technology:** the concept of a distributed shared database of transactions
- **A blockchain:** a particular implementation of a distributed ledger system that uses the same basic technical architecture as Bitcoin, but could be independent of that system
- **The Bitcoin blockchain:** the blockchain that runs the Bitcoin digital currency. In the wider world, this is often confusingly referred to simply as “the blockchain”, but we avoid that usage

Permissions

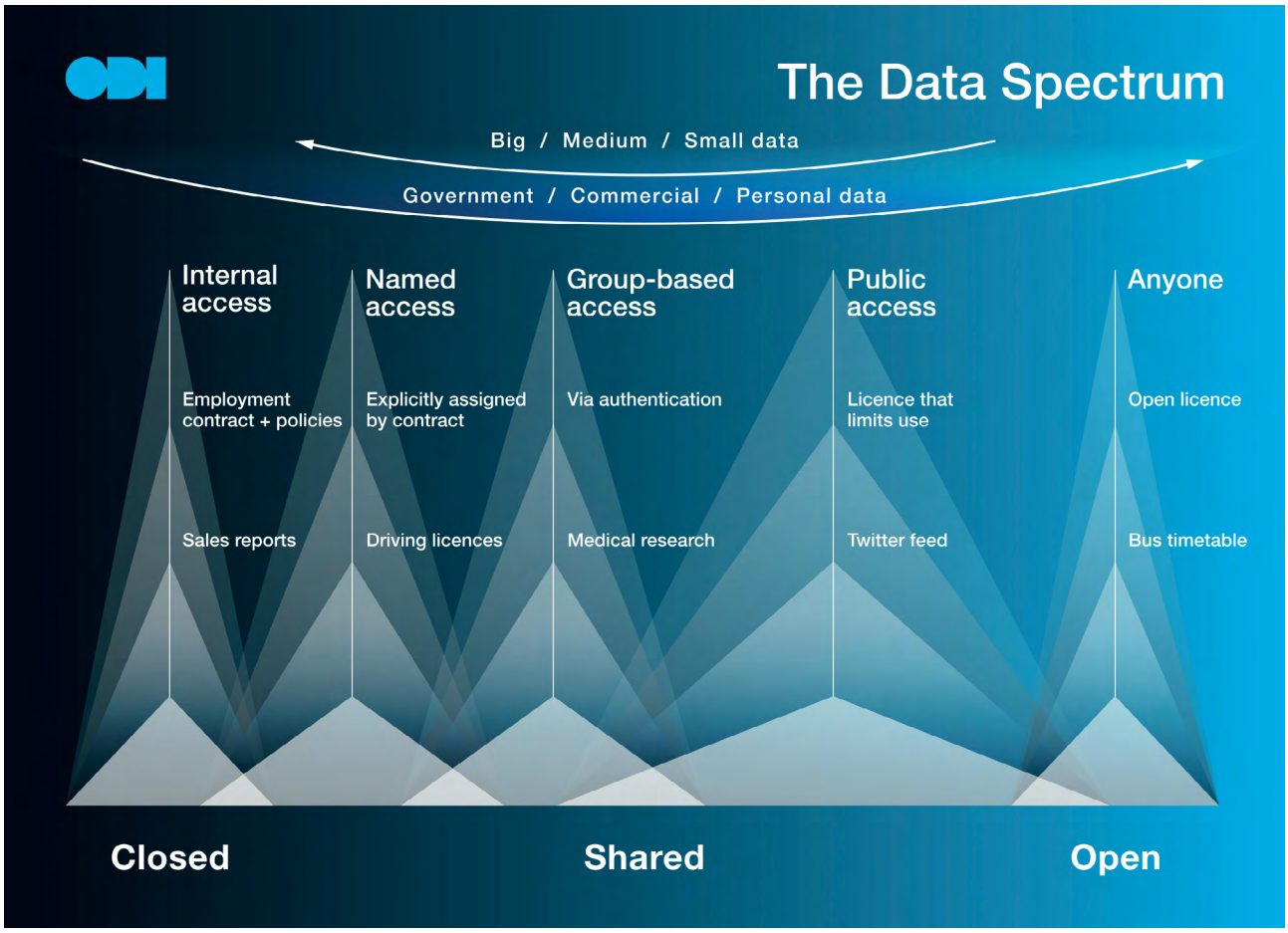
While classical blockchains (such as Bitcoin) are public and allow anyone to write to them, some distributed ledgers work differently. Distributed ledgers can be classified as:

- **Public:** anyone may have a copy of the database and anyone may write to it. This is the classic Bitcoin-style public ledger. Sometimes referred to as ‘unpermissioned’ or ‘permissionless’
- **Permissioned:** anyone may have a copy of the database, but only certain parties may write to it – an audit log for government information, for example
- **Private:** only certain authorised users have access to the database, whether for reading or writing – a blockchain used internally within an organisation’s firewall, for example

The concepts of permissioned and private blockchains are controversial in the blockchain development community, as they may reintroduce trusted intermediaries into the system and therefore undermine one of the unique aspects of the technology.

Permission types in blockchains are aligned with the ‘closed / shared / open’ components of the ODI’s Data Spectrum, which classifies data by who has access to it.[1]

1 See http://theodi.org/data-spectrum.



Closed data is only available to a single individual or organisation (similar to a private blockchain); **shared data** is shared under various permission methods with a limited group (similar to a permissioned blockchain); and **open data** is freely available to all to access, use and share (similar to a public blockchain).

The three generations of blockchain technology

Blockchain technologies and their applications generally fall into one of three generations, defined by their use and capabilities:

- **Digital currencies** (Blockchain 1.0): storing and executing transactions carried out in purely digital currencies (i.e. Bitcoin)
- **Coloured coins** (Blockchain 2.0): using small Bitcoin transactions to represent assets that are transferrable but not digitally stored on a blockchain, such as real estate or diamonds
- **Smart contracts** (Blockchain 3.0): transactions store executable code which runs when certain criteria are fulfilled, allowing complex behaviour and fully distributed applications (Dapps)

Section summary

- **Blockchains (or distributed ledgers) are a new type of database with unique properties**
- **Blockchains suit applications where multiple readers and writers need to use the database without a trusted third-party**
- **Data stored in a blockchain cannot be changed afterwards**
- **Public, permissioned and private blockchains are differentiated by who can read or write data within them**
- **Smart contract systems enable blockchains to run executable code, allowing much more complex behaviour**

The technology landscape

The recent explosion of interest in blockchains and distributed ledgers has created much speculation on the application of these technologies. While there are many ideas being brought to the fore, there seem to be few actual implementations being referenced.

As part of our research, we explored this rapidly evolving space to pull out tangible examples of tools, technologies and applications currently in operation.

Our efforts focused primarily beyond financial applications. Because of the association with Bitcoin, there has already been a great deal of research on blockchains for cryptocurrencies which we did not want to replicate. We also focused on finding applications that illustrated the use of the technology in solving real problems and examined a variety of tools that aimed to make the underlying technologies more accessible to developers.

Given the rapid growth in this field, this section only captures a snapshot of the activity that was being undertaken in early 2016. It should however serve as an illustration of the maturity of the field and a guide to potential applications.

Trends and observations

From our landscape research, we were able to distinguish a number of interesting trends and observations.

The baseline technology is unclear

Our first observation is that the baseline technology underpinning different tools and solutions for blockchains is often difficult to understand. For instance, in many cases it is hard to distinguish between the use of 'a blockchain' versus 'the Bitcoin blockchain'. This indicates that the technology stack for distributed technologies is still very fluid and in the process of being defined.

The fluidity of the technology stack presents further issues in classifying any particular solutions as either tools, technologies or applications. For instance, Bitcoin is an application of blockchain technology, but it is also being used as a platform on which other applications are built.

Emerging platforms

At the current stage of development, platform technologies have not yet fully emerged. Many blockchain systems are bespoke, and combine both platform and application functions. As the field matures, commoditised platforms that allow applications to converge on a common, underlying technology stack are likely to appear.

Some of the most interesting applications, at least from a technical perspective, are those that use multiple distributed technologies and blockchain implementations in a single service. One example of this is Alexandria,[2] a system designed to allow users to distribute digital content. It uses Bitcoin to handle payment, Florincoin[3] (an alternative cryptocurrency) to create a searchable index, and the InterPlanetary File System[4] or 'IPFS' (a distributed hash table or DHT) for content storage and distribution.

Such emerging hybrid stacks show that complex services can be entirely implemented in a shared, distributed manner. We find one promising stack to be the combination of Ethereum[5] for the application or logic layer, IPFS for bulk storage, and BigchainDB[6] for database storage. This shows the beginning of a true platform ecosystem mirroring the common LAMP[7] web stack.

2 See http://blocktech.com.
3 See http://florincoin.org.
4 See https://ipfs.io.
5 See https://ethereum.org.
6 See https://www.bigchaindb.com.
7 Linux, Apache, MySQL, and PHP.

The blockchain field is focused more on technology than real-world problems

Startups working with blockchains tend to be highly focused around technology. Many distributed technology startups begin by laying out their technical vision, rather than focusing first on an end-user need.

Many blockchain solutions are also related to other disruptive and emerging trends in the tech sector, such as the Internet of Things and the sharing economy. The focus on technology over real-world problems is a common feature in those areas as well, and can be compounded by the combination.

Existing applications of blockchain technology

As part of our research into distributed ledgers, we compiled a list of the tools, technologies and applications that we looked into.[8]

From 'marriage on the blockchain' (Woods, 2015) to 'AI gladiators' (Borah, 2015), blockchain technologies are being experimented with in many different ways. In a crowded field – where many aim to capitalise on associated hype – it can be hard to identify useful applications of blockchain and distributed technologies that might have a lasting effect.

We have identified several categories of notable applications. These are generally focused around a real-world problem that requires the unique properties of the technology. The best examples are from organisations that were committed to solving a problem before they adopted blockchains as a possible solution, as opposed to many who attempt to fit the problem to the technology.

The following sections identify the areas we think are worth watching, along with examples of organisations in the space.

Document and Intellectual Property verification

The following services are designed to prove the existence of digital documents by recording a unique identifier and timestamp in public blockchains.

- **SealX and GeniusX:** contract notary services and IP certification systems (bitproof.io)
- **Blocksign:** a document verification service (blocksign.com)
- **Factom:** a service for distributed record storage (factom.org)

8 This list is published under an open licence for anyone to access, use and share at dlt-research.labs.theodi.org.

Monitoring supply chains

The following services aim to improve supply-chain transparency and efficiency by tracking materials and products through the production process. How physical products are reliably linked to digital assets is an open question.

- **EverLedger:** stores records of diamonds to prevent insurance fraud and diamond smuggling (everledger.io)
- **Provenance:** tracking supply chain provenance (provenance.org)
- **Skuchain:** tracking supply chain provenance, especially for food (skuchain.com)

Building a peer-to-peer economy

The following services aim to automate various sharing economy transactions using smart contracts, without the need for centralised platforms.

- **Slock.it:** a service for renting, selling and sharing things that can be locked; combining IoT with blockchain (slock.it)
- **Airlock.me:** a decentralised electronic security service offering keyless access protocols for smart property (airlock.me)
- **Lazooz:** a decentralised transportation platform for ride-sharing (lazooz.org)
- **WeiFund:** a crowdfunding and equity platform based on Ethereum (weifund.io)

Governance

The following services aim to use smart contracts to inform decision-making processes, and in some cases create Decentralised Autonomous Organisations (DAOs) or even decentralised nations.

- **Colony:** a tool to help organise distributed, democratic company-like entities or organisations (colony.io)
- **Freecoin:** an alternative currency for democratic social organisations (freecoin.ch)
- **BoardRoom:** a service that organises decision-making processes (boardroom.to)
- **BitNation:** a collaborative platform for governance for organisations and even virtual nations (bitnation.co)

Digital content distribution

The following services aim to provide a means to distribute digital content equitably – whether for IP protection, or fair remuneration – in a way that does not rely on centralised platforms.

- **ALEXANDRIA:** a standard allowing users to distribute digital content (blocktech.com)
- **UjoMusic:** a service for distributing creative digital products, such as music, focused on generating greater transparency, fairness and profitability (ujomusic.com)
- **Blockai:** a digital asset distribution service (blockai.com)

False promises

While there are promising applications as identified above, a great many of the ideas out there are 'vapourware', with no viable implementation or model. For instance, development on Honduras' land registry, which is being turned into a distributed ledger by Factom, has stalled with no working system (Rizzo, 2015). This is particularly important, as this example is used repeatedly to show that blockchains can be useful in traditional government applications, but has not yet shown any results.

There are also many instances of old ideas being brought back to life with an application of new technology sheen. For instance, tracking benefit payments and how they are spent is a policy idea that has been proposed and rejected in the past, but is now reappearing with blockchains. Many such projects failed for good reasons in the past, and the addition of blockchains will not change those reasons, which is more often social or cultural than technological.

Section summary

- **The distributed ledger technology stack is still emerging, with unclear boundaries between platforms and applications**
- **Many blockchain applications are technology-centric rather than focused on user needs**
- **While there are good examples of blockchains being used to build useful services, they are currently in the minority**
- **Interested parties must be alert to blockchain hype and myths when surveying the landscape, and should focus on those applications solving real problems**

Scaling and linking blockchains

Blockchains are emerging from their origins in cryptocurrencies and being explored as a mechanism for storing data of other kinds. We are very early in our understanding of when and how best to use blockchain technology. We need to anticipate and plan for what happens when blockchains scale from low levels of use to potential ubiquity for other applications, like recording marriages (Alexander, 2014), registering land ownership (van Wirdum, 2015) and maintaining supply-chain provenance (Provenance, 2015).

Blockchains are maintained by a distributed network of nodes: computers that store the blockchains and may add data to them. There are drivers for having a few blockchains that are each maintained by a large number of nodes and for having many blockchains that are each maintained by a small number of nodes. It is likely it will end up somewhere in the middle. The suitable number of blockchains needed to support the applications that use them will change over time. We have to ensure that it is possible to adjust: to split blockchains or to merge blockchains as required, and to migrate data between them.

What are the drivers for having fewer blockchains?

Blockchains are attractive as a data store because their distributed nature makes them robust and tamper-proof. Robustness ensures that the data is always available. A blockchain being tamper-proof guarantees data integrity; even if some nodes are compromised, the other nodes will not accept changes they make to the blockchain.

A blockchain that is maintained by a single node could be struck by a hardware failure, or could rewind and rewrite the blockchain it is maintaining without detection.

Blockchains that are only maintained by small numbers of nodes can get into situations where the majority of the network is owned by a single organisation or cartel. This happened with the GHash Bitcoin mining pool in 2014 (Hern, 2014), and was the reason behind Onename's recent migration from Namecoin to the Bitcoin blockchain (Onename, 2015). When more than half a blockchain is owned by a single organisation, it is possible that they can collaborate to alter the content of the blockchain or to accept invalid transactions.

The fact that small networks of nodes undermine the utility of a blockchain is a driver towards having a few, large-scale blockchains maintained by many nodes.

What are the drivers for having more blockchains?

The size of a blockchain grows over time because it is an append-only data store: you can add data to a blockchain, but you can never remove it.

At the time of publication, the Bitcoin blockchain is around 70GB in size. It has been growing steadily by about 2.5GB/month (though that rate is increasing). With a limit of 1MB/block and one block every 10 minutes, the maximum rate of increase in size will be just over 4GB/month. While there is work underway to change this, the choice of size is controversial and the outcome unknown in the long-term.

The Bitcoin blockchain is relatively small, as transactions take up very little data. Other applications for blockchains may require more storage, or a speedier rate of growth (larger blocks and/or more frequent addition of blocks to the chain).

Every node in a blockchain network needs to be able to store and process the entirety of the blockchain. Blockchains that are used by lots of applications will be large in size. The vast majority of data within a blockchain supporting multiple applications will be irrelevant to the application any particular node is interested in. Some nodes might only be interested in land registry data, some only in statements of copyright ownership.

Large blockchains require nodes that are interested in a given application to take on all the data from other applications using the blockchain. They might hesitate not only because of size but due to ethical concerns about the data those blockchains contain. These are drivers towards having more, smaller and more specialised blockchains.

Linking blockchains

To find the balance between a few large and many small blockchains, blockchains will need to be able to split and merge over time, and for this to work, they will need to be able to refer to each other.

It is also important that new data infrastructure, built on blockchains, is compatible with the Web we have already. We will need to be able to link to items in a blockchain from outside, on the wider Internet. After all, data is at its most useful when it can be referred to and linked. Therefore, transactions on blockchains should have stable URLs and an equivalent of HTTP redirection to point to updated locations. We will need a standardisation effort to create standard URL schemes for blockchain transactions, if these systems are to link into our global data infrastructure.

Data standards and archiving

How do we standardise storage in systems so that we get a single network of data, as opposed to having to use a different storage system every time we want a new type of information? What are the data protocols for distributed storage? How do we talk about, and perhaps enforce, ownership and licensing?

Like other Internet phenomena, blockchains should be automatically archived for posterity. Whatever the archiving institution – whether the Internet Archive or something else – they will need to be able to discover the data that need archiving, automatically, and standards will be an important part of this.

Section summary

- **Fewer, large blockchains will be more secure**
- **More small blockchains will be more scalable**
- **Blockchains will need to be able to split or merge over time to balance their size and security**
- **We need to create URL and related standards to link blockchains into our global data infrastructure**

Privacy and security

The irreversibility and transparency of public blockchains mean they are probably unsuitable for personal data. We need to be careful when designing blockchain systems not to infringe on people's privacy, and to account for a world in which we have doxing, identity theft and the right to be forgotten.

The examples in this section are not necessarily mentioned because anyone has suggested providing these kinds of data in blockchains. But blockchains are being investigated for a range of purposes, across the Data Spectrum, and we need to explore their limits.

What irreversibility means for privacy

There are types of data for which the difficulty of change could lead to problems – particularly personal data.

For example, in the UK personal insolvency notices are published in the London Gazette when people are declared bankrupt. They are published in paper copies of the Gazette (which are relatively hard to get hold of or search) and published on the Web. Requests are sometimes made for notices to be removed from the website to avoid them being as easy to find, because:

- people do not want their personal insolvency to be known about (they may have a right for that event to be forgotten under the law)
- people have made a transition to a new gender; they may have a legal right for previous data to be altered to ensure it is interpreted according to their new gender
- the notice includes the current address of a person who is attempting to avoid an abuser, and thus the notice increases the risk of discovery and harm to that individual

It is easy to imagine other scenarios where blockchains could be used to hold data about people that might seem innocuous at the time but where the situation changes such that data should no longer be held in the same way. For example, a recent change in UK law means that company directors' birth dates are no longer published by Companies House.[9] Data about people applying for planning applications, holders of licences and those in public office is routinely published and could change in similar ways.

Beyond government, blockchains offer a great opportunity for people to collaboratively create datasets in a peer network without a central authority. We can see the evidence of the demand for data rating teachers, about landlords, about where sexual offenders live – all of which would have a significant privacy impact if published (Neyfahk, 2015).

9 UK Parliament (2015) Small Business, Enterprise and Employment Act, 96(3). Available at: http://www.legislation.gov.uk/ukpga/2015/26/section/96/enacted. [Accessed 2016-05-30].

Arguably, we have a situation even now where once data is published on the Web, it can never truly be removed. Certainly removing a page from Google's search results under a right to be forgotten does not actually remove it from the Web, rather it makes it harder to find. What is different about blockchains is that if a court were to attempt to legally compel the removal of data from them, it would be both hard to do and have very disruptive side-effects, which we explain in the next section.

Removing data from a blockchain

Imagine a blockchain were found to contain the names and addresses of children at risk of abuse (we deliberately pick something most people would say should not be publicly available). To clear that data out, over half the nodes that maintain the blockchain would have to work together to rebuild the blockchain since that data was added. This process is similar to rebuilding from a backup: while being rebuilt, the blockchain would be rewound to a previous state, days or weeks or even more out of date. During this time (and rebuilding blockchains deliberately takes time), the data would not be up-to-date. This might also be a time when unwanted changes to data that was trustworthy could creep in.

Alternatively, a court could try to compel the entire set of nodes to be shut down. Putting aside that nodes may reside in different legal jurisdictions, that would have huge practical implications. It would mean removing all the rest of the data held in the blockchain as well as the target of the order. Given the use of blockchains that people envisage often involves the same blockchain holding many types of data and supporting many types of applications, there is a real risk that, pragmatically, bad data simply has to continue to exist in order to prevent massive disruption to the provision of good data for other applications.

What transparency means for privacy

When new data is added to a blockchain, peers in the network check it to ensure it is valid, to avoid fraud by rogue nodes. The data that the peers need to check must be stored transparently in the blockchain.

Sometimes – as with Bitcoin – personal data is required in order to verify that a transaction in the blockchain is valid. For a node to check a Bitcoin transaction, it must have access to all previous transactions and be able to check that the person giving the Bitcoins actually has them to give. It must therefore be possible for any node to reconstruct the full financial history of every person exchanging Bitcoins: how many they have, where they got them from, and whom they spend them with.

This is intensely personal information, highly revealing of the details of someone's life; after all, would you publish all your bank statements on the Web? The only shield is the pseudo-anonymity of the Bitcoin address, which is easily breached if the address is associated with

a donate button on a blog. Those who trade Bitcoins are therefore advised to hold several addresses and not to transfer Bitcoins between those accounts to avoid others linking them together. It is not clear how many Bitcoin holders are aware of this risk, and 'security-through-obscurity' is well-known to be insufficient.

Some of the proposed uses for blockchain – such as to record auditable benefits payments – threaten to expose this kind of information about a much wider range of people, the benefits they receive and with whom they spend them.

Blockchains do not have to expose personal data directly to reveal private information about people. A blockchain recording visits to health practitioners (including midwives, mental health teams and AIDS clinics) does not need to include the entirety of someone's health records to reveal information about them. Much like phone records (Mayer & Mutchler, 2014) or browsing histories, this metadata may be sufficient to reveal personal details.

Designing privacy-protecting blockchains

There are ways to design the content of blockchains and the network that supports them, that limit the level of disclosure that they entail.

First, whereas anyone can join the Bitcoin blockchain, it is possible to use a permissioned distributed ledger as a method of resolving conflicts within a peer-to-peer network of trusted nodes. When nodes can be trusted, they can control what becomes public, and therefore hide data in the blockchain that should not be shared. The security of all the nodes in such a trusted network needs to be guaranteed as every node will have a copy of all the relevant data, and the network needs to be protected against spoofing, but, in general, if you have a trusted network many privacy issues are no more problematic than they are in centralised systems.

Second, blockchains could be used purely to provide a timestamp for information held elsewhere. Content that could require redactions in the future can be made available as usual on the Web, with transactions in the blockchain containing simply a pointer to the content and its hash. If the content needs to be taken down, the fact that the content existed at a given point in time could remain in the blockchain; the stored hash alone would not enable the reconstruction of the removed content. If the content needs to be changed, the existing hash would no longer match the content, so applications are able to detect that something has changed. If the changes are legitimate, the reasons for them could be stored in a later, overriding, transaction.

This pattern of using blockchains purely as a timestamping mechanism and not as a data store has the additional benefit of being more likely to scale in the face of large amounts of data needing to be recorded. On the other hand, it shifts the burden of robust, distributed data storage, which is one aspect of the interest in blockchains, to other protocols, whether the Web, BitTorrent, or IPFS. Projects that are focused on robust distributed storage might not need a blockchain at all.

Finally, it is possible to encrypt data stored within a blockchain. The main problem with this approach is that if the decryption key for encrypted data is ever made public, the encrypted content is readable by anyone with that key; there is no way of encrypting the data with a different key once it is embedded within a blockchain. Conversely, if the key is ever lost, the data cannot be read. And there is the problem of sharing the key for the data amongst all those who legitimately need to be able to read it.

On top of that, a well-used blockchain will be a potentially eternal datastore, and over a sufficiently long period any encryption will be broken whether by discovery of loopholes, backdoors, or the advent of new techniques such as quantum computing.

Regardless of the approach taken to designing blockchains, every blockchain contains transaction data. That data needs to be designed so that it is not disclosive in and of itself, which may be a tricky balance as that data might also be necessary to assess whether the transaction is valid and therefore prevent fraud or errors. Transactions should also be designed so that they cannot be used to add comments that might include personal data.

Blockchains are not necessarily bad for privacy; it all depends on how they are designed. As stated by Vitalik Buterin (2015): "blockchains do not solve privacy issues, and are an authenticity solution only". Anyone experimenting in the area should be thinking through the implications. As the ICO's guidance around privacy by design suggests, designers should be carrying out a privacy impact assessment or similar process up-front, to ensure that the transparency of the information stored in the blockchain does not infringe on people's privacy. Unlike with other technologies, getting it wrong is really hard to reverse.

Section summary

- **Immutable data storage in blockchains may be incompatible with legislation which requires changes to the "official truth"**
- **Once added, removing data from blockchains can be impractical and highly disruptive**
- **Even if personal data is not stored on a blockchain, metadata can be sufficient to reveal personal information**
- **Blockchains by themselves are not a solution for personal or private data**
- **Any encryption used is likely to be broken in the future**
- **Bad blockchain design decisions are very hard to reverse**

Practical experimentation

During early 2016, the ODI Labs team carried out some practical exploration of blockchain technologies and how they could be applied to data infrastructure. This was done in order to understand the technology more deeply, but also to drive out further issues through experimentation.

Creating a blockchain

The first step was to create a blockchain to experiment with. The team opted to build on Multichain,[10] a blockchain software toolkit that supports Bitcoin and other blockchains, but adds support for metadata and custom assets.

As software developers, it was reasonably simple to get a system up and running, although this would be confusing and difficult for anyone without technical experience. There is a clear need for user-friendly prototyping systems so that non-technical users (for instance policy-makers) can explore their own use cases.

Storing data on a blockchain

The team built a small open source software library[11] around the blockchain that would allow arbitrary information to be stored with any transaction. In the standard Bitcoin-style system we were using, only a small amount of information can be stored with each transaction.[12] The total amount of information that could actually be stored was only 640 bits per transaction. By comparison, a single tweet can hold up to around 2650 bits (Munroe, 2014).

One of the use cases in mind for this data storage was to be able to store a cryptographic hash of the contents of a URL. The hash is a "fingerprint" for the data pointed to by the URL; the data cannot be recovered from it, but a third party could use the hash to detect if the data has changed since the hash was created.

Auditing open data

The Government Digital Services in the UK has written about guaranteeing the integrity of a register (Potter, 2015), explaining how a distributed database could be used to check that an official register has not been tampered with.

For our exploration, we decided to follow the example proposed, and create a blockchain

10 See http://www.multichain.com.
11 Available at https://github.com/theodi/multichain-client.
12 Using OP_RETURN metadata.

version of the Food Standards Agency premises ratings dataset. If the FSA website was compromised, and ratings altered, the information on our blockchain – which could not be deleted or altered – could be used to detect the tampering. We successfully imported the full rating data (rather than just an audit hash) into the blockchain, storing the premises ID, along with the inspection date and rating.

Searching blockchains

While the team successfully stored open data in a blockchain, various issues emerged when trying to use the data.

Firstly, finding a particular record is non-trivial. To find a particular food hygiene rating for a premise, we needed to carry out a brute-force search of the blockchain. Many search systems are available, and indeed there are many sites for searching and viewing the Bitcoin blockchain, which index the blockchain into a searchable database. The problem then becomes that you are now relying on the integrity of your search index to ensure you find the right information.

If the idea of the blockchain is to prevent centralisation (due to lack of trust or other reasons), we cannot trust search indexes maintained by others. In order to find anything, not only would each blockchain client have to have the entire blockchain stored, but also store (and update) a search index built from it, which could be considerably larger. There is a need for search indexing to be built into blockchain software if it is to be used for data.

Incentivising maintenance

The test blockchain we created consisted of only two nodes, mining and sending fake currency to each other. The key to the integrity of blockchain technology is its decentralised nature. Once one person or organisation owns the majority of nodes in a blockchain network, this integrity breaks down, as discussed earlier.

If we were to open up our audit blockchain to the wider public, how would we incentivise mining? With Bitcoin, once a miner adds new transactions to the blockchain, there is a fixed reward (currently 25 BTC, almost £7,500 in today's exchange rate). As the size of the blockchain network increases, so does the computational power required to add new blocks (if using a 'proof of work' algorithm); O'Dwyer and Malone (2014) have compared the electricity usage of the Bitcoin network to that of Ireland. As a result of this, and the rewards at stake, there are entire server farms devoted to Bitcoin mining.

Currently, the test blockchain we created is tiny, and a few well-intentioned people might be happy to mine it for free. However, as the blockchain grows, mining will get harder, and the worthless currency the blockchain 'rewards' miners with won't help them cover the electricity bills the computation will end up costing them.

Section summary

- **Simple prototyping tools are required so that non-technical users can explore blockchain use cases**
- **Only a very small amount of information can be encoded with each transaction in current blockchain software (including Bitcoin)**
- **Distributed search indexes will be required for trusted search of blockchain data, increasing the space requirements**

Conclusions

Data is part of the infrastructure of the modern world. It is essential to the operation of society, and it is vital that we learn how to build, maintain and strengthen our data infrastructure. Distributed ledgers are a potentially important technology for enabling a shared data infrastructure, and are worthy of investigation.

However, new technologies always go through a hype cycle. The challenge at the beginning of that cycle is to identify the uses and applications that will stand the test of time. Blockchains are unusual in that mistakes made in early deployment could last a long time, and could cause significant damage, especially if deployed carelessly with personal data.

Blockchains could be used to build confidence in government services through public auditability. They also hold great potential for collaborative maintenance of data assets, enabling widely distributed data collection and publishing for applications such as supply-chain information. Smart contracts also have promise for the future across many application areas.

However, in our research we have seen many cases where people attempt to bolt old, failed or impossible policy and business ideas onto the new technology, or to unnecessarily reinvent things that work perfectly well. Many other cases show familiar organisational models being rebuilt as permissioned ledgers based on blockchain technologies, but this ignores the core innovation of the technology and its promised transformation.

We have seen many ideas that would put new personal data into blockchains but learnt that, if misused, this will create significant new privacy issues. The core problems that blockchain technologies help to address – of distributed maintenance by collaborating organisations – is of growing importance and an area that shows some promise, but few are considering it.
Success in data infrastructure design will come from convening sectors (such as finance,

agriculture, or healthcare), identifying common challenges and then determining which technology approaches – whether blockchains or not – are the most appropriate in helping to address them.

Blockchain technology is a new and powerful tool in our toolbox. We must use it when it is the right tool for the job at hand.

Recommendations

- **Organisations must remember to start with user needs, rather than preselecting technologies that may or not be appropriate**
- **Organisations considering blockchain technology should be aware that there are many other distributed technologies available, and it is important to assess needs properly in order to arrive at the right technology choice**
- **Organisations should convene across sectors to identify common data infrastructure needs, and then decide how those can best be met and which technologies should be used**
- **When experimenting with blockchain technologies, researchers and developers must be aware of the privacy implications of storing information in a public immutable database; what is done cannot be easily undone**

About the ODI

The Open Data Institute (ODI) connects, equips and inspires people around the world to innovate with data. It is independent, nonprofit and nonpartisan, founded in 2012 by Sir Tim Berners-Lee and Sir Nigel Shadbolt. From its headquarters in London and via its global network of startups, members and nodes, the ODI offers training, research and strategic advice for organisations looking to explore the possibilities of data.

Bibliography

Alexander, R. (2014). *The First Blockchain Wedding*. [Online] Available at: https://bitcoinmagazine.com/articles/first-blockchain-wedding-2-1412544247 [Accessed 2016-05-20].

Borah, P. (2015). *DAO Wars*. [Online] Available at: https://github.com/consensys/dao-wars [Accessed 2016-05-23].

Buterin, V. (2016). *Privacy on the Blockchain*. [Online] Available at: https://blog.ethereum.org/2016/01/15/privacy-on-the-blockchain/ [Accessed 2016-05-20].

Hern, A. (2014). *Bitcoin currency could have been destroyed by '51%' attack*. The Guardian [Online] Available at: https://www.theguardian.com/technology/2014/jun/16/bitcoin-currency-destroyed-51-attack-ghash-io [Accessed 2016-05-20].

Hearn, M. (2016). *The resolution of the Bitcoin experiment.* [Online] Available at: https://medium.com/@octskyward/the-resolution-of-the-bitcoin-experiment-dabb30201f7 [Accessed 2016-05-20].

Gartner. (1995). *Hype Cycle Research Methodology.* [Online] Available at: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp [Accessed 2016-05-20].

Greenspan, G. (2015). *Avoiding the pointless blockchain project.* [Online] Available at: http://www.multichain.com/blog/2015/11/avoiding-pointless-blockchain-project/ [Accessed 2016-05-23].

Mayer, J. & Mutchler, P. (2014). *MetaPhone: The Sensitivity of Telephone Metadata.* [Online] Available at: http://webpolicy.org/2014/03/12/metaphone-the-sensitivity-of-telephone-metadata/ [Accessed 2016-05-20].

Munroe, R. (2014). Twitter. In: *What If?* London: John Murray, p. 217–221.

Neyfahk, L. (2015). *California's Sane New Approach to Sex Offenders.* [Online] Available at: http://www.slate.com/articles/news_and_politics/crime/2015/04/california_s_sane_new_approach_to_sex_offenders_and_why_no_one_is_following.html [Accessed 2016-05-20].

O'Dwyer, K. and Malone, D. (2014). *Bitcoin Mining and its Energy Footprint.* [Online] Available at: https://karlodwyer.github.io/publications/pdf/bitcoin_KJOD_2014.pdf [Accessed 2016-05-21].

Onename, (2015). *Why Onename is Migrating to the Bitcoin Blockchain.* [Online] Available at: http://blog.onename.com/namecoin-to-bitcoin [Accessed 2016-05-20].

Provenance, (2015). *Blockchain: the solution for transparency in product supply chains.* [Online] Available at: https://www.provenance.org/whitepaper [Accessed 2016-05-20].

Resnikoff, P. (2015). *I'm Imogen Heap. And This Is Why I'm Releasing My Music on Blockchain.* [Online] Available at: http://www.digitalmusicnews.com/2015/10/05/im-imogen-heap-and-this-is-why-im-releasing-my-music-on-blockchain/ [Accessed 2016-05-20].

Rizzo, P. (2015). *Blockchain Land Title Project 'Stalls' in Honduras.* [Online] Available at: http://www.coindesk.com/debate-factom-land-title-honduras/ [Accessed 2016-05-23].

Swan, M. (2015). *Blockchain: Blueprint for a new economy.* 1st ed. O'Reilly Media.

van Wirdum, A. (2015). *Honduran Gov't to Build Land Registry Initiative on Bitcoin Blockchain.* [Online] Available at: http://cointelegraph.com/news/honduran-govt-to-build-land-registry-initiative-on-bitcoin-blockchain [Accessed 2016-05-20].

Woods, T. (2015). *This couple got married on the blockchain.* [Online] Available at: https://technical.ly/brooklyn/2015/11/11/couple-got-married-blockchain/ [Accessed 2016-05-23].