www.pipelinepub.com
Volume 18, Issue 2

Assessing Your Data Provenance Score

By: Georg Greve, Andrea Worrlein

Provenance describes the history of the state, custody, or location of something. At first, it was a term used mostly for the arts to answer questions such as, “Is this really a Picasso?” Provenance, though, also applies to the digital world, where it is part of the bedrock on which our entire digital existence rests. So, here are three aspects left largely unconsidered about provenance, and how to best assess a company’s own digital provenance.

Try proving you're NOT a dog

Just ten years ago, our world was still largely “analog-first.” Our correspondence, contracts, invoices, important documents and certificates all had analog originals. That is rapidly changing now. Most of our certificates, obligations and assets are increasingly “digital-first.” This process started years ago, but accelerated dramatically with the COVID-19 crisis, and most of us have not yet fully considered the implications.

The analog world is much more secure and resistant to fraud. Falsifying physical documents also leaves physical evidence, often requires special equipment, and sometimes even physical access to a certain location. By comparison, the digital world is ridiculously easy to falsify: any computer that is fewer than ten years old and has access to the Internet is enough to manipulate information anywhere in the world.

This fundamental difference reaches all the way down to the information that makes up who we are, who we know or what we have done. Good luck trying to prove you did not sign that contract without digital provenance. In fact, good luck trying to prove you are not a dog.

Not your keys, not your provenance

All your digital provenance is based on a technology that is quite literally built on sand—bits and pieces that are indistinguishable from one another and have no physical properties that can discern their age. Proving anything online always relies on cryptography, a peer-reviewed, human selection of universal mathematical laws.

Thanks to cryptography, we can identify data sets with near-absolute certainty and detect any changes to these data sets, no matter how minute. Cryptography also allows us to prove ownership of digital assets by virtue of signed data. Anyone holding the matching key to a signature is considered the owner. Possession may be nine-tenths of the law, but control of the keys is one hundred percent of digital ownership.

To complicate matters further, computers are agnostic to time. Data sets and keys may contain time stamps. But those only provide the time that the user set—either directly, or by changing the system clock as desired. This way, data can claim to be from the past or the future, zero Delorians required.

There are companies that will provide the service of holding an organization’s keys, just as there are organizations that will provide time-stamping services. But that also means your entire provenance is beholden to these companies. And very often their terms of service make your provenance their property. You are merely allowed to make use of your own digital provenance within their terms of service. Violate their terms of service, or find the company caught in some regulatory dispute in a country far away, and you may lose it altogether.

Digital provenance lives in gated communities

When you use Gmail, only Google really knows what is real. The same is true for any other cloud service—and if you self-host, only you know the truth behind your digital provenance, but you might find it near impossible to prove that to anyone else.

Any data shared from any such platform always comes with an implicit disclaimer of “I hereby warrant that platform provider X says this is so.” There is no good way for a third party to verify the veracity of anything without cooperation from the platform. But what if the company you are using does not consider it worth their time to help you prove your digital provenance? What if they flat out refuse because it would cost them time and money?

This also applies to professional communication and collaboration platforms. Worldwide, 300 billion e-mails are sent every day, and it is often unclear who is behind them. This makes it all the more important to ensure the verified origin of an email, which can be easily verified by any mail client or server. Messaging, ticketing or groupware are also predestined areas of application for authentic communication. Trust is a wonderful virtue, but in digital communication channels it can quickly become a boomerang.

For the majority of people on the Internet, their digital provenance is likely even more important than their most important paper documents. But they keep their provenance in a gated community, in a house they don't have the keys to, with no control over the security guards. It is nearly impossible to share critical information in an authentic way with anyone outside that gated community. And because of this, all these gated communities are a lot like Hotel California—that is, by design.

Let them eat cake?

Data sovereignty is one of the biggest challenges that humankind is facing. Europe especially has historically been very concerned about this—but has yet to deliver sustainable answers. Amid the technical community, so-called "self-hosting"—owning your servers and using open-source technology to set up your own services—has been promoted as the solution for decades. While this may be possible for people with sufficient technical background and skills to follow, to the majority, the suggestion is akin to Marie Antoinette telling her starving population to eat cake.

While companies of a certain size could opt to follow this approach at least for critical functions, very few of them do. This is because data sovereignty—like security—is considered a cost center and primarily driven by compliance. Many of these laws were put in place because the majority of business managers do not understand enough about technology to be considered literate. But while technical illiteracy might be considered a blocker for upper management, it doesn't have to be.

Far too many corporate boards lack a single member with a technical background. Instead, the cost points for compliance, namely data sovereignty and security, are delegated to the CFO. Data sovereignty requires a number of decisions about when a business can safely rely on external economies of scale, and when a company must be careful to maintain its competitive edge moving forward. These decisions are strategic in nature, and the answers will be specific to most businesses. Delegating such decisions to someone illiterate on the subject leaves many organizations with a strategic blind spot.

Defining digital provenance

For most companies it makes no commercial or strategic sense to keep everything in-house. Data sovereignty is a complex topic, with a wide range of options starting from setting up your own hardware fabrication and building your own data centers, a path that some large tech companies have opted for and that otherwise only nation-states should be considering, all the way to just signing up for the whole product line on one of the hyperscale clouds. There are always tradeoffs to be considered. For example, one could choose a secure collaboration solution but run it on a hyperscale cloud that provides platform or infrastructure as-a-service (P/IaaS).

The Swiss Academy of Engineering Sciences defines data sovereignty as␟ "the right and the ability of individuals or organizations to control and to use autonomously the data they produced or collected or that concern them."

Data sovereignty is never absolute. Any of the choices above will leave you with a different data sovereignty score. But whatever the score of your data sovereignty, there is always a separate score for your digital provenance. If we follow the definition above, data provenance should be understood as the right and the ability of individuals or organizations to prove the existence, custody, veracity and origin of data they produced or collected or that concern them.

We might also add an element of control to the disclosure of the proof to third parties, but that is already covered by data privacy. Most people understand data privacy fairly well, and data sovereignty at an intuitive level. But due to a lack of literacy in technology, they do not understand that data provenance is merely assumed, implied or promised by traditional approaches.

Even if you fabricated all your chips, owned all your data centers, and ran all your services, you would still find yourself unable to prove data provenance to anyone because they would have to trust you absolutely. And if you were to use third-party providers, they would not only have to trust you but also trust all the providers involved.

Iron-clad digital provenance, block by block

For the longest time there has been a missing link between sovereignty, privacy and provenance. While certain infrastructures—aptly named "trust centers"—existed to provide some level of third-party trust, these trust centers are expensive, have been compromised, and establish single points of failure. Bitcoin solved that.

Bitcoin was the first of the blockchains and established the world's first decentralized, linear, chronological record that is immutable, secure and robust against censorship. While most people only know it as digital currency from the financial papers, its implications reach much further. Not only did it spawn a plethora of innovation in blockchains with a wide range of approaches, blockchains also became fertile ground for groundbreaking follow-on innovation in the area of digital provenance.

Thanks to blockchain, users can have full control over their own data, store it on their own devices, hold all the keys, and still prove their digital provenance to anyone. More importantly, they can do this inexpensively, without single points of failure or a single trusted intermediary. Blockchain has made digital provenance possible in theory, but most solutions do not yet include it in practice. For some, that is an oversight. For others, it is a strategic choice to keep their users dependent on their services.

The future is self-sovereign identity (SSI)

The combination of digital sovereignty, privacy and provenance results in what has emerged as self-sovereign identity (SSI), and is the answer to the platform and surveillance economy. With SSI one is able to:

- Prove their identity and parts about themselves in a privacy-friendly way that does not require the participation of a third party;
- Control their own keys in a way that is far more convenient and far more secure than traditional passwords, without a third party;
- Secure their digital provenance across gated communities, in a way that it can be shared, verified and proven whenever, wherever and for as long as one so decides.

SSI will impact all businesses, either as a user of information technology (IT), or as a provider of services to people or companies using IT. In future, understanding your digital provenance score and understanding how to work with the digital provenance requirements of your customers, partners and suppliers will be essential.