Erschienen in: Bulletin of IEEE Technical Committee on Digital Libraries (TCDL) ; 13 (2017), 1

Using the Blockchain of Cryptocurrencies for Timestamping Digital Cultural Heritage

Bela Gipp[1], Norman Meuschke[1], Joeran Beel[2], Corinna Breitinger[1]

[1]University of Konstanz
78457 Konstanz, Germany
{first.last}@uni.kn

[2]Trinity College Dublin
Dublin 2, Ireland
joeran@beel.org

ABSTRACT

The proportion of information that is exclusively available online is continuously increasing. Unlike physical print media, online news outlets, magazines, or blogs are not immune to retrospective modification. Even significant editing of text in online news sources can easily go unnoticed. This poses a challenge to the preservation of digital cultural heritage. It is nearly impossible for regular readers to verify whether the textual content they encounter online has at one point been modified from its initial state, and at what time or to what extent the text was modified to its current version. In this paper, we propose a web-based platform that allows users to submit the URL for any web content they wish to track for changes. The system automatically creates a trusted timestamp stored in the blockchain of the cryptocurrency Bitcoin for the hash of the HTML content available at the user-specified URL. By using trusted timestamping to secure a 'snapshot' of online information as it existed at a specific time, any subsequent changes made to the content can be identified.

Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries – Collection, Standards, Systems issues. H.3.5 [Information Storage and Retrieval]: Online Information Services – Web-based services.

Keywords

Web archiving; online news preservation; trusted timestamping; blockchain applications.

1. INTRODUCTION

An increasing amount of information that contributes to humanity's cultural heritage is born digital-only. News articles, blogs or Tweets are prominent examples. Today's online information and news sources define our retrospective view on events and contribute to shaping public opinion [4, 7]. Yet, online-only content defies the traditional archiving and preservation approaches of print media.

Changing the content of physical newspapers and books is difficult and replacing them in public libraries is hardly feasible. However, changing words in an article that only appeared online is relatively simple and changes can occur at any time. The question that arises is: How can we ensure that online information remains available in an untampered form for current and future generations?

Developing strategies for securely preserving digital cultural heritage is thus a major undertaking and has been identified as a grand challenge by the German Society for Computer Science[1]. Ensuring the integrity and originality of preserved digital information is an important subtask of that challenge.

[1] https://www.gi.de/index.php?id=4174

Once information has been published online, it may be modified by its creator(s) over time for a variety of reasons, such as to:

- update the article with new information;
- remove certain details;
- change the tone or sentiment expressed in the article, e.g. from positive to negative, or vice versa;
- censor sensitive information to promote the agenda of a government, individual, or special interest group;

It is not uncommon for news articles to be revised by editors to portray a particular policy or individual in either a more favorable or unfavorable light. For example, a few hours after The New York Times published an article declaring Senator Sanders' success at the polls during the 2016 U.S. presidential elections, the article's text was edited and its tone changed from overwhelmingly positive to slightly dismissive. The article's original title "*Bernie Sanders Scored Victories for Years via Legislative Side Doors*" was modified to a substantially less celebratory title that read: "*Via Legislative Side Doors, Bernie Sanders Won Modest Victories*" [6].

These types of modifications, also called 'stealth editing', regularly occur in online news and information sources. Being able to identify and track when such changes occur can be relevant to readers. In this paper, we present an automated system to preserve, identify changes, and timestamp information from important online sources of digital cultural heritage, including news outlets, websites, blogs, or social media.

2. BACKGROUND

Trusted timestamping is a process for proving that certain digital information existed at a given point in time [3]. In a previous paper, we showed that cryptocurrencies, such as Bitcoin [5], can serve as a decentralized trusted timestamping (DTT) service if the hash value of the digital data to be timestamped is embedded into a transaction recorded in the *blockchain* of the cryptocurrency [1]. We introduced OriginStamp (www.originstamp.org) as a free service that implements this approach.

Decentralized trusted timestamping using cryptocurrencies offers the following benefits over established timestamping protocols:

1. decentralized, cryptographic integrity validation of the timestamping process;
2. high incentives for computing nodes to contribute to the decentralized record-keeping process at the heart of the Bitcoin protocol;
3. minimal effort for users: no need to setup specialized hardware or software;
4. low cost of operation, which allows us to provide the service free of charge.

For the technical description of the implementation of DTT, please refer to [1].

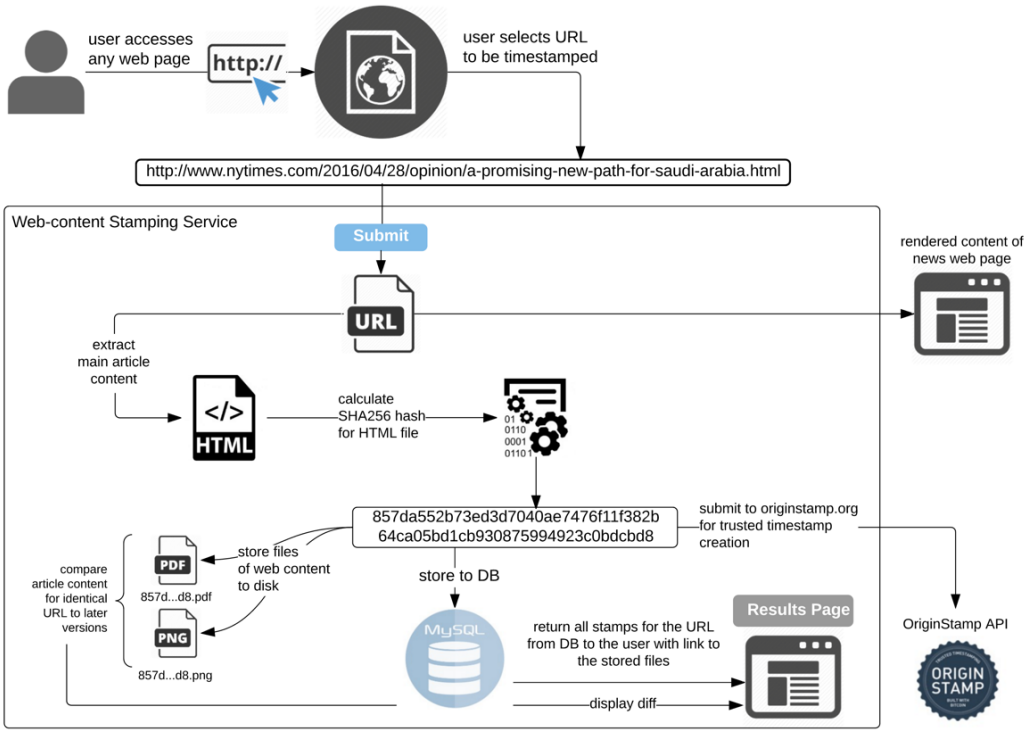

Figure 1: Overview of system functionality

3. PROTOTYPE

We present a web-based service that can be used to create automatic trusted timestamps for any online content. Figure 1 gives an overview of the functionalities of the system.

The web-interface of the system can be accessed at: www.isg.uni-konstanz.de/web-time-stamps. The code has been published under an open source licence and is available upon request.

The service focuses on the preservation of textual content and news articles, since these forms of online media are the most vulnerable to censorship and retrospective editing. However, the service can be used to timestamp any web content where the HTML and the images of the web page can be extracted. The system enables the following:

1. automatic downloading and periodic archiving of any web page (URL) specified by a user;
2. identifying textual changes and visualizing the detected modifications;
3. creating tamper-proof decentralized trusted timestamps for the downloaded webpages that make it impossible for any changes in the content to go unnoticed.

Additionally, users who wish to check for censored or blocked web content, can select from several proxies within the service to simulate accessing the same URL from different countries.

The web archiving feature is similar to the service of the Internet Archive's 'Wayback Machine' (https://archive.org/web/). However, existing archiving services [2], including the Wayback Machine, do not generate tamperproof trusted timestamps to secure the state of web content at a specific time.

Furthermore, they do not allow for a side-by-side comparison of modifications, or give users the option to be notified when changes occurred on a specific web page.

There are two other significant drawbacks of existing web archiving services. First, users must place all of their trust in the good intentions of the centrally operated archiving service. Second, the security of the storage mechanisms used by the service can never be fully guaranteed. Both of these shortcomings are solved by securing the snapshots generated by archiving sites with decentralized trusted timestamps using the blockchain.

The prototype lets users specify the frequency with which timestamps should be created for the content residing at the specified URL (see Figure 2). For example, a specific blog page could be timestamped every day or once a week. Each time a new timestamp is generated, the system automatically generated a diff of the content (see Figure 4). The user is notified via email if any changes occurred.

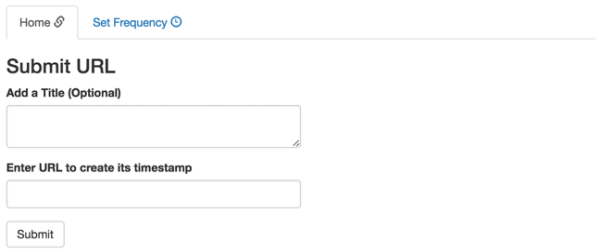

Figure 2: URL submission interface. Once the user submits the URL a timestamp of the current state of the digital content residing at that URL is automatically generated. The content is stored in PDF and PNG form for preservation.

Figure 3 shows the system interface for browsing user-submitted URLs. For each URL, the unique SHA256 hash generated from the content snapshot, as well as a PDF, a PNG, and a link for verifying the trusted timestamp (which currently redirects to originstamp.org) are displayed.

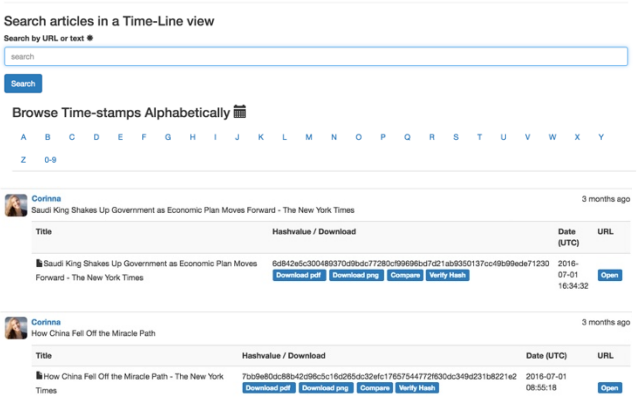

Figure 3: User-submitted online content overview

We encourage readers to examine the text modifications shown in Figure 4 for themselves. The figure visualizes retrospective modifications made to a New York Times article titled "Tension Rising, Saudi Monarch Ousts Ministers". The original article, published May 8th, 2016 is compared against the version available online on May 15th, 2016. Newly added text passages are highlighted in green, while deleted text is marked in red.

"They are clearly meant to help in terms of implementing the vision going forward."None of the decrees addressed the country's political structure, and missing from Prince Mohammed's vision were any moves toward allowing Saudis a role in choosing their rulers. None of the newly appointed officials were women.

Other changes were in line ~~forward."Other changes also jibed~~ with goals articulated in the new vision.The ~~vision, showing the tight coordination between Prince Mohammed, who has emerged as the country's most dynamic official despite being second in line for the throne, and his father, King Salman, who maintains ultimate authority.The~~ Ministry of Hajj, an important body in a country that derives much of its international standing ~~legitimacy~~ from its management of Islam's holiest sites, was changed to the Ministry of Hajj and Umra. While the Hajj pilgrimage happens once a year, the Umra pilgrimage can be done year round, ~~year-round,~~ and Prince Mohammed has spoken of raising ~~expanding~~ the number of Umra visitors as a source of unexploited income. The body's ~~its~~ minister was also replaced.Many ~~people~~ among Saudi Arabia's large population of young people ~~youth population~~ — more than two-thirds of the country's 20 million citizens are under 30 — speculated on social media, often with sarcasm, ~~media~~ about the duties of the newly created General Authority for Entertainment. ~~Commission for Entertainment. It is~~ a surprising body in a hyperconservative country where public movie theaters are banned and many people flee abroad for vacations and weekends.Some Twitter users responded by posting photos of members of the religious police smashing musical instruments under the hashtag #The Entertainment Authority in Arabic."Oh Lord!" one Twitter user wrote. "Theater, opera, cinema and stadiums for women."Another wrote: "They split the electricity ministry and the water ministries for fear of electrocution. Joke courtesy of #The Entertainment Authority."Prince ~~spend their vacations and weekends abroad. Prince~~ Mohammed has spoken about the importance of providing Saudis with more ways to enjoy life in their country, although it remains to be seen what new options will be provided under the new plan.

Follow Ben Hubbard on Twitter @NYTBen.Clifford Krauss reported ~~provided.~~

~~Clifford Krauss contributed reporting~~ from Houston.

A version of this list appears in print on May 8, 2016, on page A1 of the New York edition with the headline: Tension Rising, Saudi Monarch Ousts Ministers. Order Reprints| Today's Paper|Subscribe

Figure 4: Visualization of changes made to a New York Times online article that was timestamped and tracked for subsequent changes by our system

4. OUTLOOK

To increase the impact of our service, we plan to develop plugins for browsers and popular web applications, such as Twitter, Facebook, and Reddit to simplify and speed up the process of timestamping and tracking web content for end users.

Additionally, we are currently realizing an API that enables developers of other services and applications to integrate our service for timestamping and tracking digital content at scale.

To increase the trustworthiness of timestamped web content snapshots, we are working on a distributed version of the service. As explained in Section 3, our service currently stores snapshots of the monitored online media content on a user's computer and submits the hash of the corresponding files to OriginStamp. This approach exhibits an inherent weakness, because the user might alter the files that represent the snapshot of the web resource and submit the hash of the altered files to OriginStamp. Hence, establishing a definitive chain of evidence that a web resource existed exactly in the version timestamped by a user requires redundant timestamping of the resource by multiple parties independently of each other.

We plan to overcome this weakness by distributing the timestamping and monitoring process to multiple nodes. When a user submits a URL for tracking, the service will automatically perform the timestamping and tracking on the user's computer and on additional nodes participating in the service. The nodes will routinely compare their results to identify discrepancies. If the timestamped snapshots differ among nodes, the service can trigger notifications or log the event for future reference.

Distributing the service to multiple nodes will also facilitate proving situations in which web resources present different content to different users. For example, online news sites might present different content to users in different countries. Checking for censored content using a world map is possible in the current prototype, yet the process is limited to a single machine. The future version of our service will automatically distribute the timestamping and tracking process to nodes in different countries. Routinely performed checks of the snapshots taken by the different nodes can easily prove differences in content.

5. AKNOWLEDGEMENTS

We would like to thank W. Detho and S. Schneider for their contributions to the development of the prototype.

6. REFERENCES

[1] Gipp, B. et al. 2015. Decentralized Trusted Timestamping using the Crypto Currency Bitcoin. *Proceedings of the iConference 2015* (Newport Beach, California, Mar. 2015).

[2] Gomes, D. et al. 2011. A survey on web archiving initiatives. *International Conference on Theory and Practice of Digital Libraries* (2011), 408–420.

[3] Haber, S. and Stornetta, W.S. 1990. How to Time-Stamp a Digital Document. *Advances in Cryptology—CRYPTO '90 Proceedings*. 3, 2 (1990), 99–111.

[4] Kim Fridkin Kahn, P.J.K. 2002. The Slant of the News: How Editorial Endorsements Influence Campaign Coverage and Citizens' Views of Candidates. *The American Political Science Review*. 96, 2 (2002), 381–394.

[5] Nakamoto, S. 2008. Bitcoin: A peer-to-peer electronic cash system. (2008). Available: https://bitcoin.org/bitcoin.pdf

[6] Sullivan, M. 2016. Were Changes to Sanders Article "Stealth Editing"? *The New York Times*.

[7] Wanta, W. et al. 2004. Agenda setting and international news: Media influence on public perceptions of foreign nations. *Journalism & Mass Communication Quarterly*. 81, 2 (2004), 364–377.