Ockr Specification

Bormann, Fabian
bormann.fabian@gmail.com

Ockr GitHub Community
Contributors

November 25, 2023

1 Abstract

Ockr aims to set a standard for creating machine-readable and reliable documents, enabling the verification of their authenticity. In short Ockr tries to tackle three problems:

1. How can we make sure that a document is machine-readable?
2. How can we ensure that a document has not been modified?
3. How can we ensure that a document really has been issued by a certain authority?

2 Motivation

Most of the documents today are created in a digital form and are shared online. The file itself serves as a vehicle for the actual information. File hashes are often used to prove the authenticity of the information. However, the file hash can be changed over time, while the information remains the same. Compression techniques are used behind the scenes while sharing the files online, as well as storing the file in cloud environments. Finally printing the file and scanning it again will result in a different file hash. Images or documents that store the information in images are not machine readable. While process automation became crutual in many industries, many processes still rely on manual work because of the lack of machine readable documents. OCR (Optical Character Recognition) is a technique that can be used to extract the information from images, but it is not reliable enough to be used in critical processes as it can not guarantee the authenticity of the information. Ockr aims to solve these problems by providing a standard for creating machine-readable and reliable information.

3 Process

To decouple the information from the file, the raw information needs to be provided and a hash of the information needs to be calculated. This hash will be used to verify the authenticity of the information and stored within a QR code on the document. Besides the hash, the QR code will also contain metadata helping a machine to re-create the information.

3.1 Document creation

As most of the documents are created digitally, the raw text can be extracted from the file such as DOCX, XLSX, PDF and other formats. In case there is no raw text available, OCR can be used to extract the text from the document but a human would need to verify the text.

3.2 OCR verification

There is no internet connection needed and no additional service required to verify the ORC result. Therefore data pipelines can ensure to have a reliable OCR result in critical processes. An algorithm breaks down the document in different parts and verifies the OCR result of each part. The algorithm can be implemented in any programming language and can be used in any environment. Besides the main hash of the information a hash of each part will be calculated and stored within the QR code. There are chars that are often misinterpreted by OCR like 0 and O. With having the hash of each part, different combinations of chars can be tried to find the correct result.

3.3 Guarantee authenticity

An open source backend service will be publicly available to register an information hash on a blockchain by a certain authority. The authority can be a person or an organization. The service will be available as a REST API. As the hash is registered on a blockchain, it would not match anymore if the information or QR code would be maliciously changed.

3.4 Authorship verification

The authority that registered the information would also need a registration on the blockchain. Anyone can register a new authority but to give a certain authority more credibility, the authority can be verified by other authorities.

3.5 Trust Trees

The trust tree is a graph of authorities that trust each other. An algorithm will be provided to calculate the trust tree and its score. The score can be used to determine the credibility of an authority.