IIR

Internet
Infrastructure
Review

Aug. 2024
Vol. 63

Periodic Observation Report

Protecting Our Customers from Ever More Sophisticated Cyberattacks

Focused Research (1)

W3C Standardization: RDF Dataset Canonicalization

Focused Research (2)

IIJ's DRM Initiatives

IIJ
Internet Initiative Japan

Internet Infrastructure Review

August 2024 Vol.63

Executive Summary

On November 30, 2022, OpenAI unveiled ChatGPT, and its capabilities sent shockwaves around the world. Various functionality has since been added to ChatGPT, and with other companies also announcing their own generative AI offerings, we're in the midst of a global boom in generative AI. The uptake of generative AI continues to advance, with many organizations using it to enhance value-added and improve efficiency.

On May 13, 2024, some 18 months after ChatGPT, OpenAI released its latest model, GPT-4o. If you've already used it, you no doubt have a feel for how it has evolved, but the many articles and videos out there introducing and reviewing GPT-4o also offer a glimpse of just how much it has changed. The speed with which generative AI technology is developing is truly astonishing.

Yet many negative aspects of AI are also being called out. In Japan this year, we have started to see an increasing number of news stories about AI being used to facilitate fraud. The incidence of deep fake images and audio of celebrities being used is also on the rise. And the potential for AI to be used in elections and propaganda has long been a concern. With important elections on the slate around the world this year, heightened vigilance will be crucial.

Against this backdrop, the European Union (EU) approved a law for regulating AI on May 21. The law will become fully applicable in 2026, and will impose four levels of restrictions on AI systems commensurate with the magnitude of the risk. High fines are to be imposed for high-risk violations. The strictest prohibitions apply to applications of AI such as social scoring and predictive policing, which can disadvantage specific individuals and groups. The next level is high-risk AI systems, which include those used in educational admissions, employment and hiring, biometric authentication, and infrastructure operations, with these systems being subject to strict conditions and obligations. Images and audio created by generative AI must also be clearly labeled as artificially generated.

The EU's AI regulation law could become the global standard for AI regulation. Understanding and complying with legal regulations is a given when using AI in business, and it is crucial that companies hold themselves to high ethical standards. As AI spreads to every corner of society, it will become increasingly important for individuals to acquire AI literacy.

The IIR introduces the wide range of technology that IIJ researches and develops, comprising periodic observation reports that provide an outline of various data IIJ obtains through the daily operation of services, as well as focused research examining specific areas of technology.

Our periodic observation report in Chapter 1 discusses messaging. Email is an important application that has been in use since the advent of the Internet. The history of email, which made it easy to send messages to a large number of recipients, is also a history of mail system administrators combating abuse. The article discusses the abuse landscape of recent years and new efforts by IIJ in this area. It also discusses developments and challenges over the past year, which has seen a major shift in the area of sender authentication.

The focused research report in Chapter 2 looks at RDF Dataset Canonicalization. RDF (Resource Description Framework) is a framework for representing information on the web and is standardized by the W3C. The article's author is involved in the standardization of RDF Dataset Canonicalization (a mechanism for canonicalizing data represented using RDF) at the W3C. The article starts by giving an overview of RDF, and then discusses the standardization effort, looking at why canonicalization is necessary, the procedures involved, and challenges faced.

The focused research report in Chapter 3 covers DRM (digital rights management) in the area of video delivery. Technologies that protect content rights are essential when distributing digital content that is easily copied. The article focused on video DRM, and it is fair to say that DRM has been a major contributor to the great popularity of today's Internet-based video delivery services. I hope this gives you some idea of what sort of processing is being performed behind the scenes when end users enjoy video content.

Through activities such as these, IIJ strives to improve and develop its services on a daily basis while maintaining the stability of the Internet. We will continue to provide a variety of services and solutions that our customers can take full advantage of as infrastructure for their corporate activities.

Junichi Shimagami

Mr. Shimagami is a Senior Executive Officer and the CTO of IIJ. His interest in the Internet led to him joining IIJ in September 1996. After engaging in the design and construction of the A-Bone Asia region network spearheaded by IIJ, as well as IIJ's backbone network, he was put in charge of IIJ network services. Since 2015, he has been responsible for network, cloud, and security technology across the board as CTO. In April 2017, he became chairman of the Telecom Services Association of Japan's MVNO Council, stepping down from that post in May 2023. In June 2021, he also became a vice-chairman of the association.

Protecting Our Customers from Ever More Sophisticated Cyberattacks

1.1 A New Era for Email

A year has passed since our last report on this topic[*1], and the email industry saw huge changes in 2023.

In the first half of this article, we report on the latest attack methodology observed by IIJ and discuss the new countermeasures we have started taking against these threats. In the second half, we report on the dramatic changes in sender authentication technology DMARC compliance rates observed over the past year.

Email infrastructure is crucial in providing organizations a means of both internal and external communication, but it is difficult to make changes once that infrastructure is built. But with attack methods and security trends ever changing, organizations face a constant need to take countermeasures. Why not take this opportunity to review your email infrastructure?

1.2 Protecting Customers from Threats

1.2.1 What is Abuse Protection?

Email services are constantly being abused as a means of sending phishing (fraudulent) emails.

In general, most ISPs (Internet service providers) and hosted email services use the combination of an email account user ID and password (credentials) for SMTP authentication, only allowing users to send emails if that authentication is successful. This authentication is used to identify users and to protect the email service from unauthorized use by third parties.

Malicious actors, however, are always stealing user credentials by some means or another and using email services to send phishing emails (account hijacking)[*2]. IIJ is not alone here. This activity occurs at other ISPs and on other companies' services, and this sort of unwelcome and fraudulent behavior on the Internet is commonly referred to as abuse.

1.2.2 Effects of Abuse

What happens when malicious actors exploit email services to send phishing emails?

In recent years, instead of simply sending unwanted advertising emails (spam), malicious actors have turned to phishing emails as a way of stealing the IDs and passwords for web services and apps from their victims, the recipients of phishing emails. Their ultimate goal is one of financial gain—stealing IDs and passwords from users duped by phishing emails enables them to then steal bank account details and credit card numbers. With the use of cloud services becoming increasingly more prevalent in recent years, this sort of activity is becoming more and more prominent.

Naturally, most email services prohibit users from sending phishing emails under their terms of use. Malicious actors are attempting to increase their attack success rates by sending large numbers of phishing emails in a very short period of time before their ability to send emails is restricted due to terms of use violations, in what is a truly shotgun approach.

*1 IIR Vol. 59 (https://www.iij.ad.jp/en/dev/iir/059.html).

*2 Some time ago, there were cases of passwords being discovered via brute-force credential attacks, but this itself is abuse and an inefficient method. In almost all cases these days, phishing emails are sent out successfully on the first try without any prior authentication attempts, so it is natural to assume that malicious actors are using some means of obtaining credentials in advance.

1.2.3 Problems of Abuse

Leaving this situation unchecked not only exposes the targeted users to harm but also adversely impacts on email services in the following ways.

- When malicious actors use email services to send out large volumes of phishing emails, this can overload the IT equipment, leading to service disruptions and reduced availability ((1) and (2) in Figure 1).
- The transmission of phishing emails results in the email service being recorded as a phishing email source by destination email servers, security vendors, etc., such that emails from other legitimate users are identified as spam and blocked at those destinations ((3) and (4) in Figure 1).
- The impact of this can be long-lived since some security vendors draw on threat intelligence from other security vendors, such that it takes time for threat intelligence to be removed from everywhere it has been recorded ((6) and (7) in Figure 1).

Hence, to ensure stability and to prevent other customers from being adversely impacted on IIJ's email service, IIJ Secure MX Service, we immediately investigate any cases of abuse and work around the clock to protect our equipment by, for example, forcibly changing email service user credentials and blocking certain communications.

1.2.4 Abuse and Secrecy of Communications

In Japan, Article 4 of the Telecommunications Business Act prohibits actions such as revealing or obtaining telecommunications handled by telecommunications carriers[*3*4]. However, when abuse has been explicitly recognized and it is highly likely that, if it is left unchecked, service users will become complicit in illegal acts that infringe on the rights of others or become

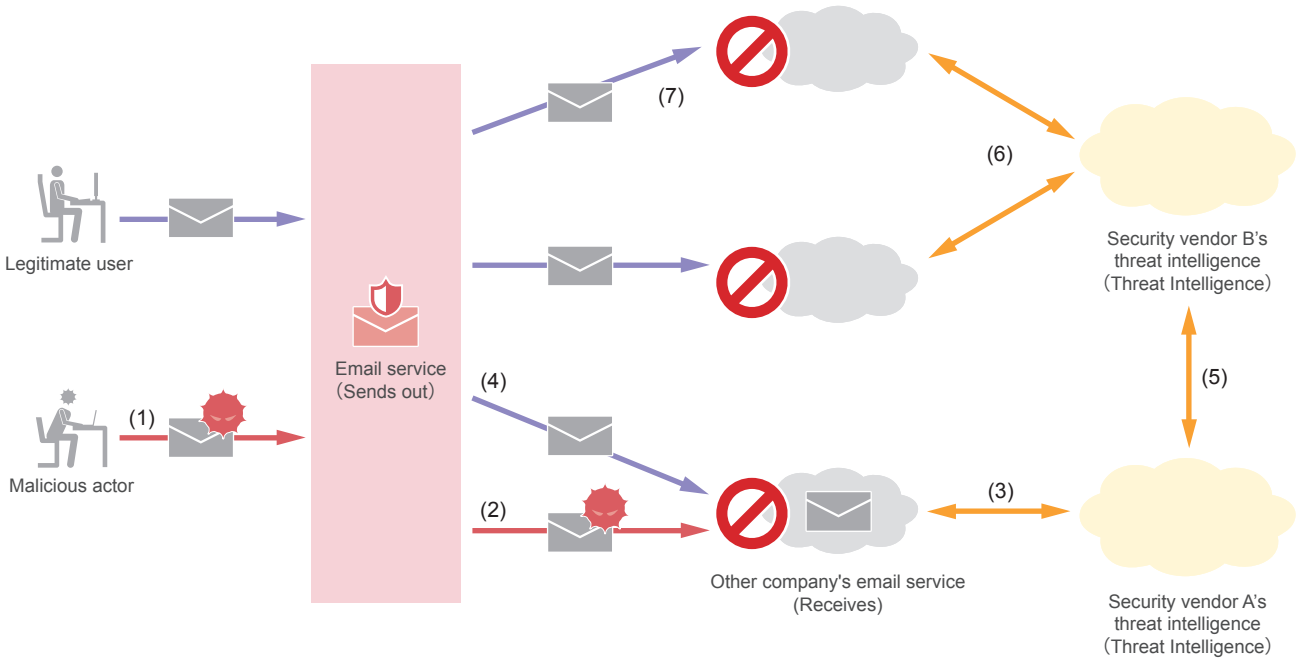


Figure 1: Adverse Impact of Abuse on Email Services

*3 This concept of secrecy of communications originally referred to postal correspondence. It is the right to prevent third parties from knowing when, by who, and with whom communications are taking place, and what the content of those communications is. In Japan, this also applies to the Internet, but countries that deal with it in this way are in the minority worldwide, and censorship is considered to be legal in the vast majority of countries. "Only four countries, including Japan, do not monitor or interfere with the Internet" (Yasuhiko Taniwaki, Kyoyo to shite to no Internet [The Internet as Culture, in Japanese], Nikkei BP, 2023; Freedom House (US-based NPO) survey, https://freedomhouse.org/sites/default/files/2022-10/FOTN2022Digital.pdf).

*4 A postal worker must look at the front of a postcard to know which recipient's mailbox to put it in, and similarly, on the Internet, communications cannot be delivered unless the IP packet headers, and in the case of email the content of the SMTP protocol content, are seen. Hence, such actions are categorized as those that, while violating the secrecy of communications, are still legitimate business activities.

victims themselves, the illegality of those (normally prohibited) actions involving telecommunications, when taken to prevent such outcomes, can be waived on the basis that they fall into the category of emergency measures or legitimate business activities.

IIJ's agreements with its users prohibit any acts that constitute abuse, and as a party to these agreements, IIJ has the ability to take various measures to deal with clear violations of the agreements.

1.2.5 Discovery of Preparations for Abuse

In the past few years, we have, through our daily operations, discovered multiple instances of test mailouts whereby someone, rather than sending phishing emails out all of a sudden, sends out a number of seemingly harmless emails a few days beforehand. The following is an example of the type of information included in the email subject line:

Email address; login ID; password; SMTP server name; port number; number sent; auth. method
Example:
`iij-taro@example.jp;iij-taro;password;mail.securemx.jp;465;2;LOGIN`

The email addresses in the recipient field of such emails are thought to be collecting the results of reconnaissance activities, and a causal link has become clear in that actual phishing emails are sent out a few days later (Figure 2).

At this preparatory stage, however, we cannot really say that these actions violate the rights of others or even that there is a high likelihood of such a violation, so we cannot necessarily label this as abuse. So even though we were aware of this preparatory stage of events, we had no basis for taking specific action until abuse actually occurred. This was a very frustrating situation for us as operators of equipment that is supposed to protect our customers while maintaining quality of service.

1.2.6 IIJ's New Initiatives

As it was, we were hamstrung and unable to protect our customers. And so we knew we needed to put a new framework in place. Bringing in IIJ's support and legal departments, we set about designing a framework for restricting communications to the extent necessary before phishing emails were actually sent out whenever we detected these sorts of preparations for the improper use of our services.

Table 1 describes the benefits of taking action before phishing emails are sent out.

After much discussion with everyone involved, we decided to implement this into our agreements through the following steps.

- Provide all customers with a detailed explanation of our new initiatives in advance.
- Also make changes that incorporate specific provisions into the IIJ Secure MX Service terms and conditions so that customers are fully aware of the changes.

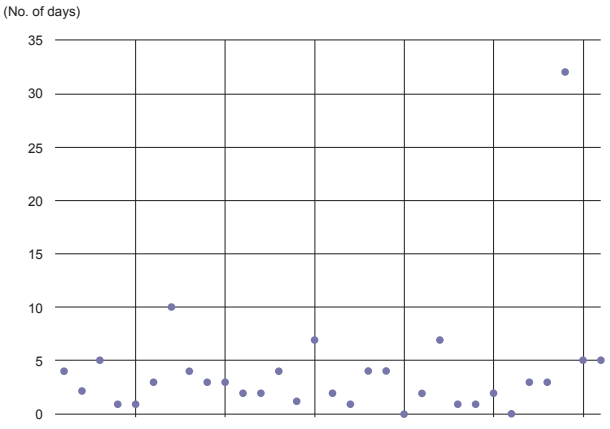


Figure 2: No. of Days After Reconnaissance that Actual Abuse Occurred (During a Particular Period)

Table 1: Benefits of Countermeasures

From perspective of	Benefits
Email service operator (IIJ)	・Can take countermeasures before phishing emails are sent, preventing service disruptions and avoiding a situation in which emails do not reach their destinations
Users	・Can detect credential breaches early ・Can limit the damage caused by email interceptions and information breaches

We have sent notifications to our existing customers' administrators, so please take a look. The relevant terms of service were included in the May 1, 2024 revision. Please refer to Article 12 (Dealing with the risk of misuse, etc.) of the IIJ Secure MX Service Individual Regulations.

Incidentally, while we refer to action taken against abuse that has already happened as our abuse response, when we detect preparations for improper use and take action to protect our customers in advance, we call this defensive action.

1.2.7 IIJ's Defensive Action

We initially discovered these reconnaissance activities in the course of our daily operations, but there is only so much we can do manually. So we have now harnessed illumino[*5], a large-scale log analysis platform deployed internally at IIJ, to use machine learning to detect events likely to constitute preparations for improper use of our services (Figure 3)[*6].

We actually did not start using Splunk for defense purposes; we were originally looking at using it to conduct investigations that would help streamline our abuse response. But it was in the course of these investigations that we uncovered these abuse preparations. We wondered whether we could use machine learning to detect this sort of activity as well, and our efforts to improve accuracy in this regard resulted in us being able to detect such preparatory actions with a fairly high probability.

1.2.8 Conclusion

The fact that the telecommunications companies through which people's communications and data pass are heavily regulated in terms of how they conduct their business is not that well known among ordinary consumers. Yet the Internet connects the entire world together. When we take action to protect our customers from malicious actors, these actions are always accompanied by efforts to protect the secrecy of communications, and we are mindful of striking a balance between these two objectives in the course of our operations.

At IIJ, we will continue working to protect our customers from the ever more sophisticated cyberattacks they face.

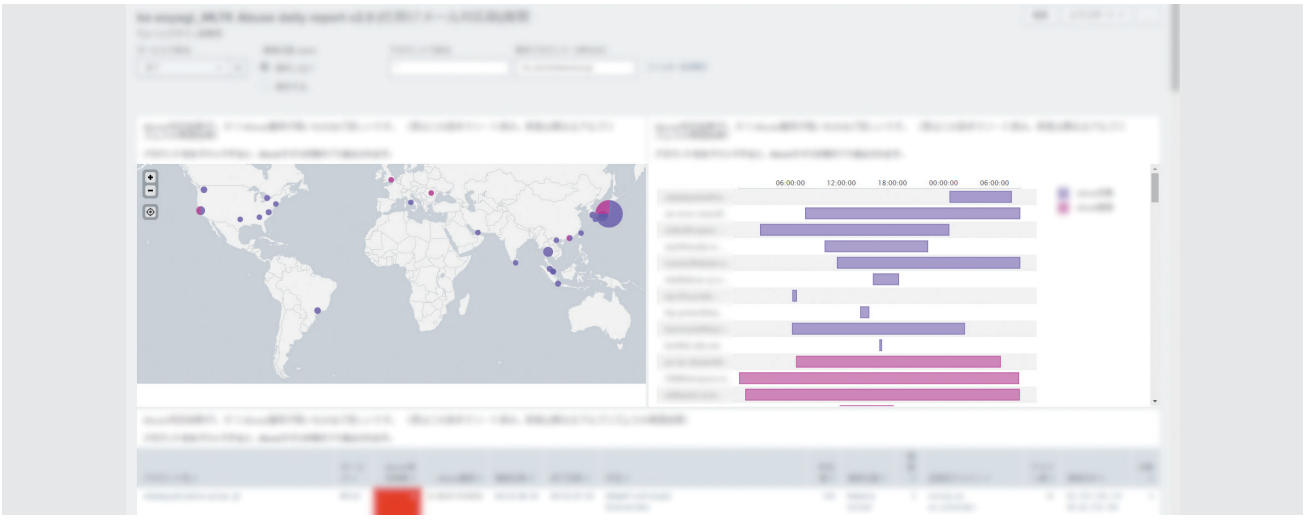


Figure 3: Splunk Dashboard for Defensive Action

*5 "illumino—IIJ's Internal Data Analytics Platform", Internet Infrastructure Review Vol. 57 (https://www.iij.ad.jp/en/dev/iir/057.html).
*6 For examples of how we use Splunk, see "Japanese Text Analysis Using Splunk", Internet Infrastructure Review Vol. 48 (https://www.iij.ad.jp/en/dev/iir/048.html).

1.3 The Big Push for Sender Authentication

1.3.1 Calls for DMARC Support in the Financial Industry and Related Developments

In February 2023, Japan's Ministry of Internal Affairs and Communications, National Police Agency, and Ministry of Economy, Trade and Industry called on financial institutions such as credit card companies to implement DMARC policies as a means of combating email spoofing[*7].

Credit card companies and the like have long been increasingly plagued by the damage caused by email spoofing, with observers exclaiming the need for countermeasures, and the official call to action seems to have set off an earnest push to take steps in that direction. This is evident from the increase in the DMARC compliance rate for financial industry domains shown in Figure 4[*8].

Only around 20% of domains had published DMARC policies as of January 2023, but one year later in January 2024, that figure had increased to 80%. Yet, many domains that have published DMARC policies still have them declared with p=none. For DMARC to be properly effective, they need to change this to p=quarantine or p=reject. A major move toward this happened in the financial industry in 2023, but a look at Japanese domains as a whole reveals that many companies are yet to make this change (Figure 5)[*9].

We will continue to focus on developments in this area in the hopes that other industries will follow the financial industry in implementing DMARC policies.

1.3.2 Google and Yahoo in the US Unveil Policy of Blocking Emails With No Sender Authentication

In October 2023, Google and Yahoo in the US announced that from February 2024 they would be blocking emails that do not support sender authentication. Both Google and Yahoo are plagued by huge volumes of spam and bulk mail every day, and to block such emails from coming into

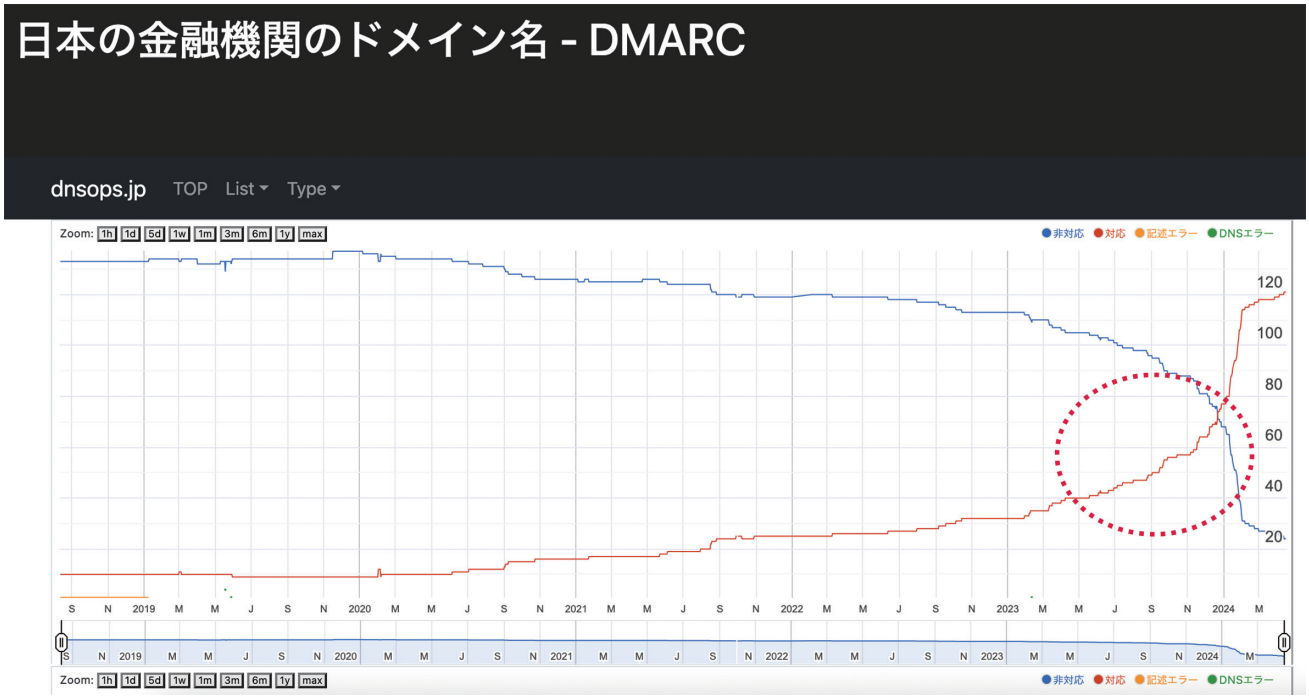


Figure 4: DMARC Compliance on Japanese Financial Institution Domains

*7 Ministry of Internal Affairs and Communications, "Call for Credit Card Companies etc. to Bolster Phishing Countermeasures" (https://www.soumu.go.jp/menu_news/s-news/01kiban18_01000184.html, in Japanese).

*8 Domain status of Japanese banks - DMARC (https://stats.dnsops.jp/chart/jp-bank/dmarc).

*9 Domain status of Japanese organizations - DMARC (https://stats.dnsops.jp/chart/all/dmarc).

their servers, the companies moved to block all emails that do not support sender authentication.

Sender authentication technologies include SPF, published in 2006 (RFC 4408), DKIM, published in 2007 (RFC 4871), and DMARC, published in 2014 (RFC 7208). In 2023, nine years after its release in 2014, DMARC was still not all that widely adopted (Figure 5)[*10*11].

The revelation that global heavyweights Google and Yahoo, which handle some of the biggest email volumes in the world, would be adopting a "no auth, no entry" policy shocked IT providers around the world. Email delivery rates are a key service indicator for mass email senders, so the move created an impetus for them to adopt DMARC with all due haste.

In the immediate wake of this, M^{3}AAWG (Messaging Anti-Abuse Working Group), which works to combat spam globally, hosted an emergency Q&A session on the announcements with representatives from Google and Yahoo at its October 2023 meeting. The session was intense, with participants, predominantly hailing from email senders around the world, asking, for instance, what level of commitment would be required and whether emails would really be blocked if they did not comply.

Subsequently, at the November 2023 meeting of JPAAWG, a working group that discusses Internet security in Japan, the matter was taken up predominantly by operators of email businesses in Japan. Of course, it is not just bulk senders but also domain owners in the form of companies and mailbox providers that need to respond to the move, and we at IIJ had also been working on a response for IIJ's our own business and personal services.

We Our email services already had a system ready for sender authentication technologies (SPF, DKIM, and DMARC) IIJ's

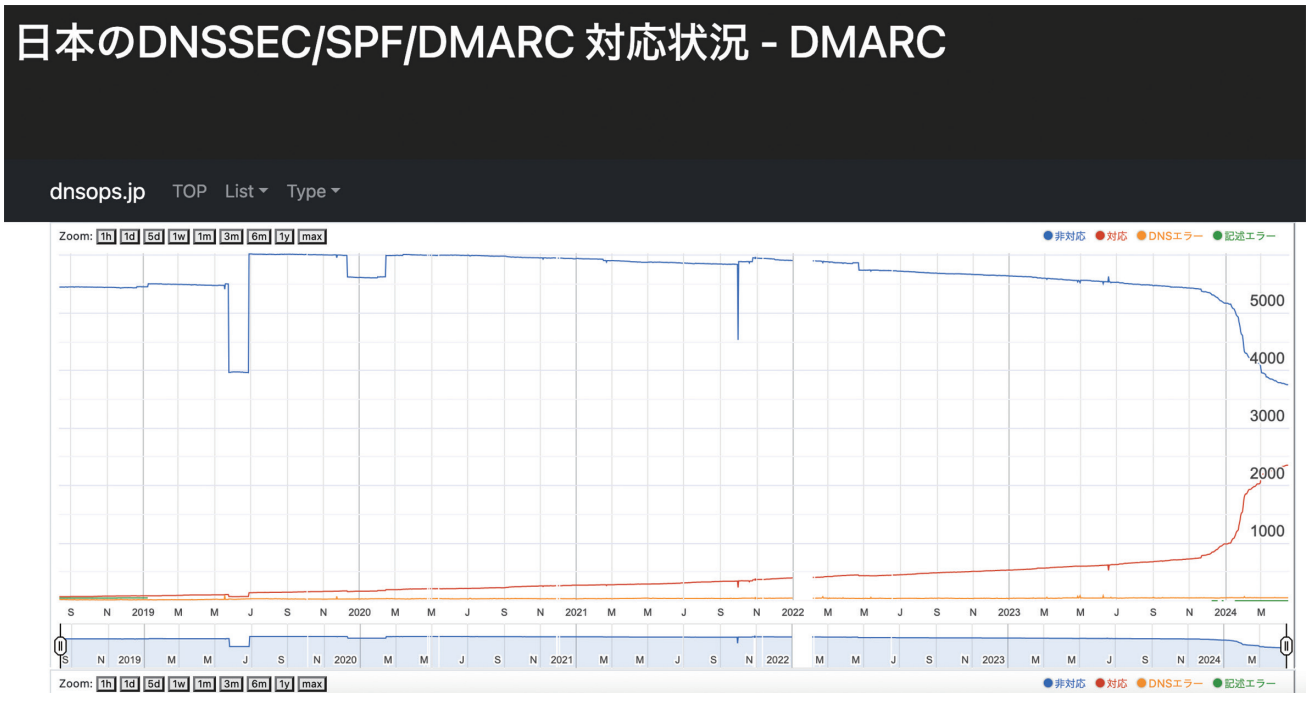


Figure 5: DMARC on Japanese Domains

*10 Domain status of Japanese organizations - DKIM (https://stats.dnsops.jp/chart/all/dkim).
*11 Domain status of Japanese organizations - DMARC (https://stats.dnsops.jp/chart/all/dmarc).

services for business customers, but customers they needed to make configuration changes and take certain steps themselves to get the features working, so we experienced a huge increase in customer inquiries about sender authentication at the end of 2023.

The efforts of the various businesses, corporations, and organizations resulted in a dramatic change in sender authentication compliance rates for emails received on the IIJ Secure MX Service (Figure 6).

The proportion of emails with DKIM signatures increased by just over 15%, and the proportion of sender domains with DMARC records (those other than none in the DMARC pie chart) increased by over 30%pt, from 42% to 75%. Given the 2022 figure was 32%, it looks like the number of organizations that have implemented DMARC records has increased for the reasons discussed above.

That said, all this does is confirm that DMARC records exist. We have not looked at whether the DMARC policies use p=none, p=quarantine, or p=reject. Domain owners should change from p=none to p=quarantine or p=reject to keep tabs on DMARC reports.

1.3.3 Problems with Sender Authentication Technologies

With the use of cloud services rising in recent years, the incidence of companies sending emails from on-premises equipment directly out onto the Internet is in decline. And so some have begun to argue that an assessment of SPF records alone is insufficient to ensure email trustworthiness.

Also, in order to send emails from multiple cloud services, some domains have been observed to exceed the limit of 10 DNS lookups when resolving SPF records, causing the SPF record lookup itself to return an error.

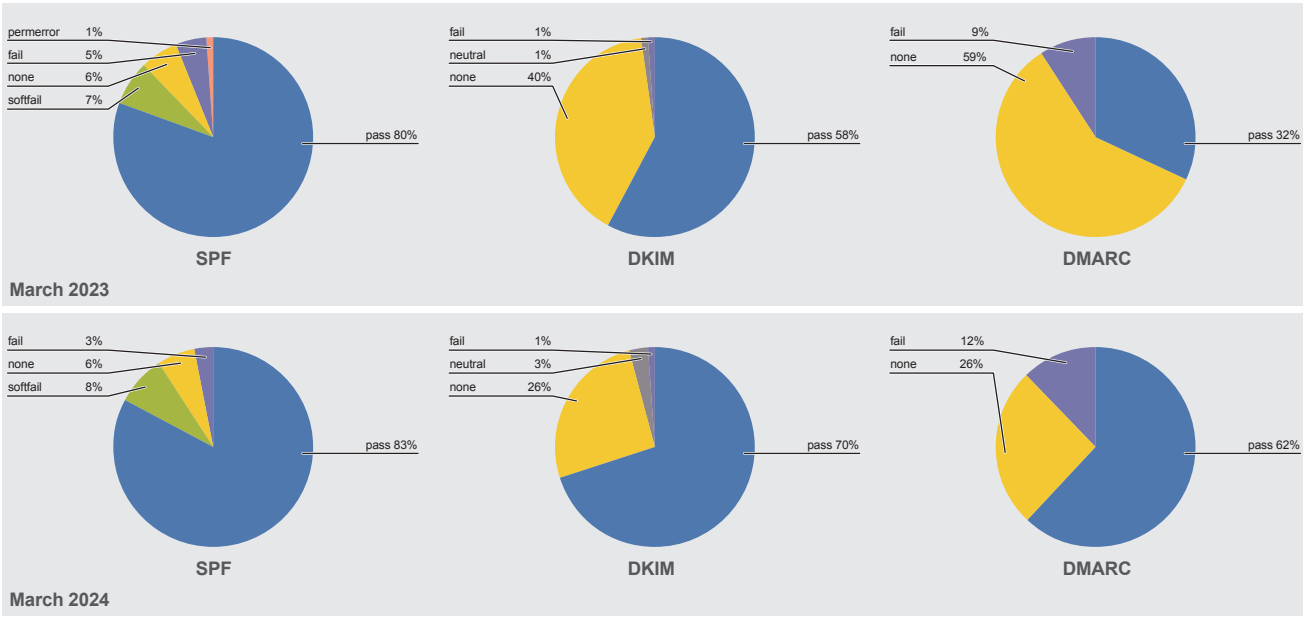


Figure 6: Sender Authentication Compliance Rates for Emails Received on Secure MX (2023 vs. 2024)

To avoid this, some providers offer services that use CNAME records to bundle the records pointed to via the SPF include mechanism into a single record. The original reason for the SPF include limit is that SPF records with many inclusions could act as DNS lookup amplifiers[*12].

Some cloud services specify SPF records with a huge number of IP address ranges, and when using your own domain name to send email from these types of cloud services, you can avoid the SPF limits by implementing DKIM signatures so that emails pass DKIM sender authentication.

Even so, DKIM is not a panacea, and you need to take action against DKIM replay attacks and properly manage DKIM key expiration terms[*13].

When it comes to DMARC too, service providers have a deep history of dealing with DKIM signature validation failures caused by long-established email mechanisms whereby headers are rewritten after emails are DKIM signed, as can happen with forwarded emails and mailing lists, for example.

ARC is designed to avoid these sorts of DKIM validation failures by re-signing emails after headers have been changed, but as with DKIM, it is left up to the receiving servers to decide which signing domains to trust.

Many issues remain to be addressed in the area of sender authentication, and IIJ is committed to tackling them by collecting and disseminating crucial information, participating in the development of IETF standards, and so forth[*14]. Efforts to support SPF, DKIM, and DMARC sparked by the recent Google and Yahoo announcements are just beginning, and it's crucial to remember that ongoing effort will be needed with respect to all of these issues.

1.1 A New Era for Email, 1.2 Protecting Customers from Threats
Isamu Koga
Manager, Mail Service Management Section, Application Service Department, Network Division, IIJ
Mr. Koga joined IIJ in 2007. He is engaged in the operation of email services and investigates email-related trends in the wild. To keep customers' email boxes safe, he serves as a communicator and public speaker on the latest attack methods, trends in spam, and countermeasures. He is also involved in a wide range of community activities, including M^{3}AAWG, WIDE Project, and openSUSE.

1.3 The Big Push for Sender Authentication
Yusuke Imamura
Lead Engineer, Mail Service Management Section, Application Service Department, Network Division, IIJ
Mr. Imamura joined IIJ in 2015. He is engaged in the operation of email services. His past experience working at IIJ Europe benefits him in fulfilling his global role.

*12 IETF Datatracker, 11. Security Considerations, 11.1. Processing Limits (https://datatracker.ietf.org/doc/html/rfc7208#section-11.1).
*13 IETF, DKIM Replay Problem Statement (https://www.ietf.org/archive/id/draft-ietf-dkim-replay-problem-00.html).
*14 IETF Datatracker, The Authenticated Received Chain (ARC) Protocol (https://datatracker.ietf.org/doc/html/rfc8617).

W3C Standardization: RDF Dataset Canonicalization

2.1 Introduction

In this article, I describe RDF Dataset Canonicalization[*1], which became a World Wide Web Consortium (W3C) recommendation in May 2024. I was involved in the standardization process at W3C. RDF Dataset Canonicalization is an algorithm for canonicalizing (i.e., normalizing or generating a serial canonicalization of) data represented using the Resource Description Framework (RDF). I explain what RDF is, what the process of RDF canonicalization entails, and when it is needed. I also go over the path we took to standardization at W3C and describe the canonicalization procedure in some detail.

2.2 What is RDF?

RDF is a W3C standard framework for describing information (resources) on the web. RDF makes it easy to link data between different databases and applications. It is widely used in areas such as the life sciences, pharmacology, and libraries for this reason. The first version became a W3C recommendation in 1999, and RDF 1.1[*2] became a recommendation in 2004. As of this writing (May 2024), work on the RDF 1.2[*3] standard is underway.

RDF represents information with three elements: a subject, a predicate, and an object. A set of these three elements is called an RDF triple. By way of example, the following is an RDF triple on the classic Japanese literary work *The Pillow Book (Makura no Soshi*, rendered below as "Makuranosoushi" in keeping with the Japan Search records) obtained from Japan Search and slightly modified for this article[*4], a site that lets you search through a wide range of Japanese content.

- Subject: <https://jpsearch.go.jp/data/bibnl-20853658>
- Predicate: <https://www.w3.org/2000/01/rdf-schema#label>
- Object: "Makuranosoushi"

RDF triples can be read like a normal sentence, as "the predicate of the subject is the object." That is, the RDF triple here can be read as "the label of bibnl-20853658 is Makuranosoushi." Here, the subject https://jpsearch.go.jp/data/bibnl-20853658 is an identifier assigned to a book by Japan Search. The predicate <https://www.w3.org/2000/01/rdf-schema#label> is a term defined in the W3C RDF Schema[*5], and the object that follows it, "Makuranosoushi", is the label (brief description) of the subject. Hence, this RDF triple indicates that the information with identifier <https://jpsearch.go.jp/data/bibnl-2085658> takes the label "Makuranosoushi".

So, RDF triples represent a lot of information using URLs[*6] like <https://jpsearch.go.jp/data/bibnl-20853658>[*7]. URLs are used so as to accurately convey the information that the data creator is trying to represent. If the subject and predicate were expressed without a URL as simply "20853658" and "label", readers would find it difficult to correctly understand where the 20853658 identifier comes from and what the meaning of the predicate label is.

RDF triples can also be drawn as a diagram with two nodes (information contained in ovals or boxes) connected by an arrow, as in Figure 1. For ease of reading, part of the URL in Figure 1, https://jpsearch.go.jp/data/, is abbreviated to "data:". Similarly, http://www.w3.org/2000/01/rdf-schema# is replaced by "rdfs:". I use this abbreviated notation below as well.

A collection of RDF triples is called an RDF graph. Retrieving additional RDF triples on The Pillow Book from Japan Search enables us to create an RDF graph like that in Figure 2.

Figure 1: Example of an RDF Triple Referring to Makuranosoushi

*1 Dave Longley, Gregg Kellogg, Dan Yamamoto: RDF Dataset Canonicalization. W3C Recommendation, May 21, 2024 (https://www.w3.org/TR/rdfcanon/).
*2 Richard Cyganiak, David Wood, Markus Lanthaler: RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation, February 25, 2014 (https://www.w3.org/TR/rdf11- concepts/).
*3 Olaf Hartig, Pierre-Antoine Champin, Gregg Kellogg, Andy Seaborne: RDF 1.2 Concepts and Abstract Syntax. W3C Working Draft, May 2, 2024 (https://www.w3.org/TR/2024/WD-rdf12-concepts-20240502/).
*4 Japan Search (https://jpsearch.go.jp/).
*5 Dan Brickley, R.V. Guha: RDF Schema 1.1. W3C Recommendation, February 25, 2014 (https://www.w3.org/TR/rdf11-schema/).
*6 To be precise, an Internationalized Resource Identifier (IRI), which is a generalization of a URL, is used.
*7 In this example, the object is expressed as a string rather than as a URL, but it is common for triples to have a URL as the object.

In this RDF graph, the label for data:bibnl-20853658 is "Makuranosoushi", and we can also see that Sei shounagon was involved in the book's production and that Moriya Shinsuke was involved in its translation.

As mentioned, a collection of RDF triples forms an RDF graph. Additionally, a collection of RDF graphs is called an RDF dataset. RDF Dataset Canonicalization is, as the name suggests, a method of canonicalizing RDF datasets. For simplicity, however, I will not distinguish between RDF graphs and RDF datasets in this article.

2.3 Blank RDF Nodes

In the example in Figure 2, nodes with the strange names _:b152539105 and _:b152573899 appear. These are a special type of node called blank nodes and do not have an identifier (URL). When creating large RDF graphs, for example, it can be cumbersome to assign URLs to all nodes. So blank nodes without URLs are sometimes used for intermediate nodes not connected to other graphs.

Names given to blank nodes are only temporary. Within the same RDF graph, the names of blank nodes may change depending on the system or environment in which they are handled. For example, while the RDF graph in Figure 3 has _:hoge and _:fuga instead of Figure 2's _:b152539105 and _:b152573899, respectively, it is treated as being the same as the RDF graph before those name replacements (more precisely, it is isomorphic to that graph).

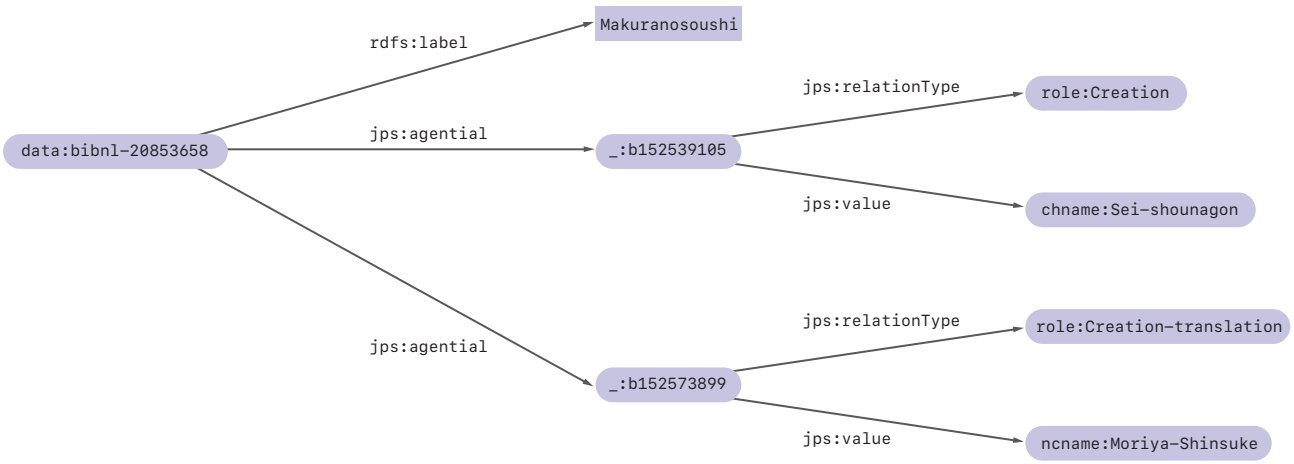


Figure 2: An RDF Graph for *The Pillow Book*

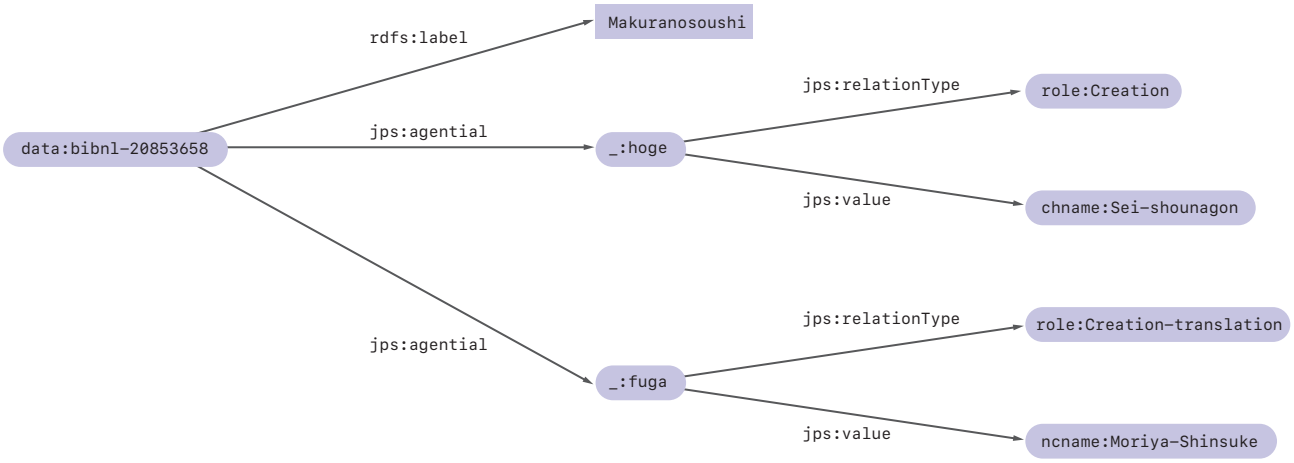


Figure 3: Another RDF Graph for *The Pillow Book*

An advantage of this is that RDF graph creators need not worry about naming blank nodes. Another advantage is that these names can be omitted when converting the RDF graph into data. For example, the RDF graph in Figure 3 can be expressed as follows using the JSON-LD[*8] specification. Here, there is no need to worry about the names of blank nodes.

```
{
    "@context": { ... },
    "@id": "data:bibnl-20853658",
    "rdfs:label": "Makuranosoushi",
    "jps:agential": [
        {
            "jps:relationType": "role:Creation",
            "jps:value": "chname:Sei-shounagon"
        },
        {
            "jps:relationType": "role:Creation-translation",
            "jps:value": "ncname:Moriya-Shinsuke"
        }
    ]
}
```

2.4 Canonicalization

Blank nodes are useful, but their lack of a fixed name can sometimes cause problems. For example, handling blank nodes can be problematic when you want to check whether two RDF graphs are isomorphic, determine the differences between graphs, or determine whether an RDF graph has been updated. Also, if an RDF graph is digitally signed by its creator, verification will fail if the names of blank nodes when the graph is signed differ from the names of blank nodes when the verification attempt is made, but due to the nature of blank nodes, it is not possible to guarantee that the names will be the same.

We therefore needed a method of assigning fixed names to blank nodes that would be independent of the system or environment. That's where RDF Dataset Canonicalization, the subject of this article, comes in. When canonicalized, the two RDF graphs in Figures 2 and 3, for example, are converted into the same graph, which is shown in Figure 4.

Once canonicalized, the blank nodes take the new names _:c14n0 and _:c14n1[*9]. These names are calculated using a predetermined method based on the URLs and strings that appear in the graph and the structure of the graph, and they are not influenced by the values originally assigned to the blank nodes, i.e., _:b152539105, _:hoge, and so forth. Thus, the same graph can be

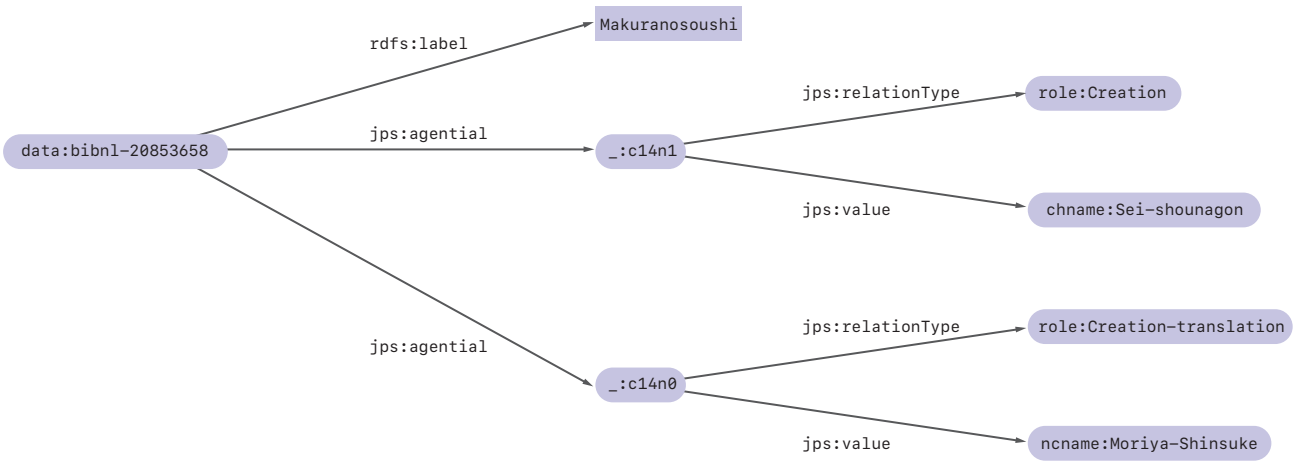


Figure 4: Example of a Canonicalized RDF Graph

*8 Gregg Kellogg, Pierre-Antoine Champin, Dave Longley: JSON-LD 1.1. W3C Recommendation, July 16, 2020 (https://www.w3.org/TR/json-ld11/).

*9 "c14n" is an abbreviation of "canonicalization".

obtained regardless of what names were assigned to the original blank nodes.

Once the blank node names are determined, a representation of the dataset called its canonical n-quads form[*10] can be generated, which, in our case, gives the canonicalized data shown in Figure 5. Using the canonicalized data, it is easy to calculate RDF graph differences, check for updates, and calculate digital signatures and hashes.

W3C Verifiable Credentials, a form of digital credentials that we covered in IIR Vol.52 (https://www.iij.ad.jp/en/dev/iir/052.html)[*11], are RDF datasets with a digital signature. Canonicalizing blank node names using RDF Dataset Canonicalization before applying the digital signature guarantees that the data that is signed will be the same as the data that is verified.

2.5 The Standardization Effort

Standardization is typically a lengthy process. The standardization of RDF Dataset Canonicalization took over a decade. Although the discussion around it started early on, one reason it took so long is that it took ages to arrive at a consensus on the need for standardization and what the optimal method would be[*12].

Discussions on the canonicalization specification first began at W3C from 2009 to 2010. In 2012, Dave Longley and Manu Sporny of Digital Bazaar proposed the Universal RDF Graph Normalization Algorithm (URGNA2012). This was followed three years later by a revised version, the Universal RDF Dataset Normalization Algorithm (URDNA2015), which became the basis for the now standardized specification.

The discussion around verifiable credentials subsequently ramped up at W3C, and 2021 saw the proposed formation of the Linked Data Signatures Working Group to work on methods of digitally signing RDF data. This effort was terminated, however, after it failed to reach consensus on the standardization of the overall signing process. The RDF Dataset Canonicalization and Hash Working Group (RCH WG)[*13], focused on RDF canonicalization, was proposed as an alternative and approved in July 2022.

And so the work to make RDF Canonicalization a W3C Recommendation finally began. On May 21, 2024, this standardization effort reached its goal of producing a W3C Recommendation[*14].

Following an invitation from the WG chair, I joined RCH WG as an Invited Expert in August 2022, and I also served as an Editor from November that year. The invitation was prompted by an article on verifiable credentials written by me and colleagues for an international conference[*15], which caught the interest of the WG co-chair.

```
<https://jpsearch.go.jp/data/bibnl-20853658> <https://jpsearch.go.jp/term/property#agential> _:c14n0 .
<https://jpsearch.go.jp/data/bibnl-20853658> <https://jpsearch.go.jp/term/property#agential> _:c14n1 .
<https://jpsearch.go.jp/data/bibnl-20853658> <rdfs:label>"Makuranosoushi".
_:c14n0 <https://jpsearch.go.jp/term/property#relationType> <https://jpsearch.go.jp/term/role/Creation-translation>.
_:c14n0 <https://jpsearch.go.jp/term/property#value> <https://jpsearch.go.jp/entity/ncname/Moriya-Shinsuke>.
_:c14n1 <https://jpsearch.go.jp/term/property#relationType> <https://jpsearch.go.jp/term/role/Creation>.
_:c14n1 <https://jpsearch.go.jp/term/property#value> <https://jpsearch.go.jp/entity/chname/Sei-Shounagon>.
```

Figure 5: The Result of Canonicalization (several URLs modified for translation purposes in this article)

*10 Generally, data in n-quads form does not need to have been sorted, and there is no limit on the number of whitespace or line feed characters used as delimiters, but here we use data sorted in lexicographical order and limit the number of delimiter characters to one. This is called the canonical n-quads form.

*11 Internet Infrastructure Review Vol. 52, "2. Focused Research (1): Verifiable Credentials and BBS+ Signatures" (https://www.iij.ad.jp/en/dev/iir/052.html). The LD Canonicalization referred to in Vol. 52 is the former name of RDF Dataset Canonicalization discussed here.

*12 Email by Phil Archer (https://lists.w3.org/Archives/Public/semantic-web/2024May/0030.html).

*13 W3C RDF Dataset Canonicalization and Hash Working Group (https://www.w3.org/groups/wg/rch/).

*14 RDF Dataset Canonicalization and Hash Working Group Charter (https://w3c.github.io/rch-wg-charter/).

*15 Dan Yamamoto, Yuji Suga, Kazue Sako, Formalising Linked-Data based Verifiable Credentials for Selective Disclosure. 2022 IEEE European Symposium on Security andPrivacy Workshops (EuroS&PW) (https://doi.org/10.1109/EuroSPW55150.2022.00013).

RCH WG mainly conducts its activities via GitHub discussions and fortnightly conference calls. Anyone can raise issues and give proposed solutions via GitHub, while on the conference calls, working group members engage in discussion to resolve issues and reach a consensus. The results of all this go through some editing on GitHub and then eventually end up in the specifications. This was my first involvement in standardization, and while I found it hard to keep up with the expert discussion, I made an effort to contribute in any way I could: proposing wordings, reviewing pull requests, and providing reference implementations, and so on.

It is crucial that W3C specifications are created in such a way that readers are able to correctly implement the content. As of this of writing (May 2024), nine open-source implementations of RDF Dataset Canonicalization have been released, having been developed in a wide range of languages: C++, Elixir, Java, JavaScript, Ruby, Rust, and TypeScript[*16]. I have also released an open-source implementation in Rust[*17].

2.6 Canonicalization Procedure

The algorithm defined in the RDF Dataset Canonicalization specification is named RDF Canonicalization algorithm version 1.0, commonly known as RDFC-1.0. Here, I give an overview of RDFC-1.0.

RDFC-1.0 consists of two steps: canonicalization, in which blank nodes in the input RDF graph are labeled, and serialization, in which the canonical n-quads form of the canonicalized RDF graph is generated.

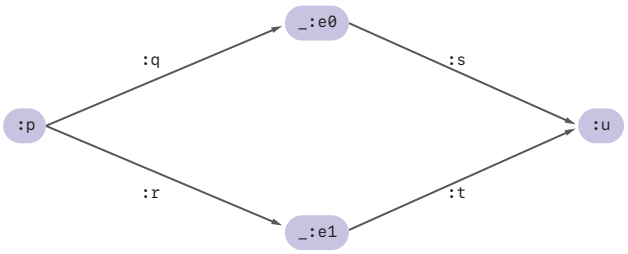


Figure 6: Example of an RDF Graph with Two Blank Nodes

In the canonicalization step, a value called the first degree hash is calculated for each blank node in the graph. This is done by passing the information around the blank node into a special function called a hash function to obtain a fixed-length block of data called the hash value. Intuitively, this constitutes giving a name to blank nodes using the information surrounding it.

If the first degree hashes assigned to the blank nodes are all different, then they can be sorted in lexicographical order[*18] so as to assign an order to the blank nodes. Labeling the blank nodes in this order—i.e., _ :c14n0, _ :c14n1, _ :c14n2, and so on—completes the canonicalization process.

I will now explain this process in detail using the example in Figure 6. This RDF graph contains four nodes, two of which (:p and :u) are normal nodes that have URLs, and the remaining two (_ :e0 and _ :e1) are blank nodes.

Extracting just the RDF triple containing blank node _ :e0 and representing it in canonical n-quads form yields this:

```
:p :q _:e0 .
_:e0 :s :u .
```

This corresponds to the information surrounding _ :e0. Here, we replace the blank node's "temporary" name of e0 with a, which yields the following string:

```
:p :q _:a .
_:a :s :u .
```

*16 Gregg Kellogg: RDF Dataset Canonicalization and Hash 1.0 Processor Conformance (https://w3c.github.io/rdf-canon/reports/).
*17 zkp-ld/rdf-canon (https://github.com/zkp-ld/rdf-canon).
*18 To be precise, they are sorted in Unicode code point order.

This is passed into the hash function, and the resulting hexadecimal bit string, 21d1dd5ba21f3dee9d76c0c00c260fa6f5d5d65315099e553026f4828d0dc77a, is used as the first degree hash of blank node _ :e0. Information about _ :e0 is embedded in this first degree hash value, and it can be used to distinguish _ :e0 from other blank nodes.

Similarly, extracting the RDF triple containing _ :e1 gives,

```
:p :r _:e1 .
_:e1 :t :u .
```

and with e1 replaced by a, as before,

```
:p :r _:a .
_:a :t :u .
```

and we then generate a first degree hash value for _ :e1 of 6fa0b9bdb376852b5743ff39ca4cbf7ea14d34966b2828478fbf222e7c764473.

When sorted in lexicographical order, the first degree hash of _ :e0, which starts with 2, comes before the first degree hash of _ :e1, which starts with 6. So we have been able to determine an ordering for e0 and e1. We then follow this order and assign the canonicalization identifier _ :c14n0 to _ :e0 and _ :c14n1 to _ :e1, completing the canonicalization process.

Crucially, the canonicalization result is always the same, regardless of what names are given to the blank nodes before canonicalization. Indeed, using the example in Figure 6, we can check that the first degree hash values do not change even if we replace _ :e0 with _ :hoge and _ :e1 with _ :fuga. As part of the process of calculating the first degree hash values, all blank node names are replaced by a, and this means that the end result is independent of the names originally given to the blank nodes.

RDF graphs that are not too complicated, such as the example in Figure 6, can be canonicalized by calculating just the first degree hashes, and this is relatively easy to do. But depending on the RDF graph, you can end up assigning the same first degree hash to different blank nodes. The graph in Figure 7, for example, contains blank nodes surrounded by information that is exactly the same, and in such situations, the nodes will be assigned the same first degree hash.

Looking at _ :e0 and _ :e1, we can see that both are objects reached from subject :p via predicate :q, and furthermore, that both are subjects that lead to a blank node object via predicate :p. Because of this, their first degree hash values will be exactly the same.

In RDFC-1.0, therefore, n-degree hashes are calculated as the next viable identification method only in cases of blank nodes being assigned the same first degree hash. The process by which n-degree hashes are calculated in RDFC-1.0 is complicated, so I will not explain it here. The interested reader is directed to the specification.

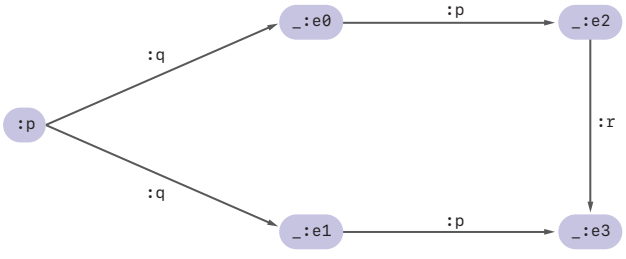


Figure 7: Example of a More Complicated RDF Graph

2.7 Canonicalization Challenges and Solutions

As should now be evident, RDF Dataset Canonicalization is, in essence, a way of obtaining canonicalized names by ordering blank nodes. So to canonicalize an RDF graph that does not contain any blank nodes, there is no need to calculate any first degree or n-degree hashes; all you have to do is perform the simple (serialization) process of representing the RDF graph in n-quads form and then sorting.

There is a misconception that RDF Dataset Canonicalization involves unnecessarily complicated processing, but the complexity of canonicalization depends on the number of blank nodes in the input RDF graph and the structure of the graph if it contains blank nodes. In practice, the process can be completed quickly with just the simple first degree hash value calculations.

Even so, there are RDF graphs with special structures containing many blank nodes for which it can take an extremely long time to calculate the n-degree hashes[*19].

For this reason, RDFC-1.0 implementations are required to place an upper limit on the number of n-degree hash calculations and terminate the process prematurely with an error if the limit is exceeded.

You also need to keep in mind that if an RDF graph contains personal data or confidential information, it may be possible to partially infer what that information is based on the canonicalization results. Since the canonicalization calculations are based on data in the graph, the canonicalized names (_:c14n0 etc.) will "partially" contain information from the graph.

Consider, for example, the RDF graph in Figure 8 showing that Alice's spouse is Bob. Say this graph is canonicalized, digitally signed, and saved as shown in Figure 9.

Say that at some point, for whatever reason, Alice wants to communicate that she is married while keeping her spouse's name hidden. Using the selective disclosure mechanism in Verifiable Credentials, she can hide her

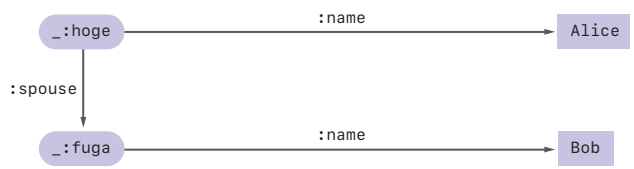


Figure 8: RDF Graph about Alice and Bob

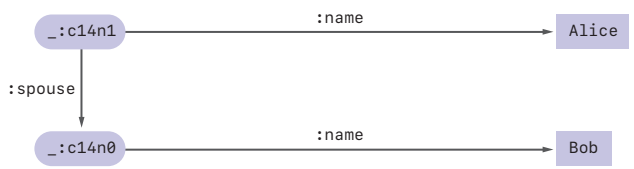


Figure 9: Canonicalized Form of FIgure 8

*19 The specification refers to these as poison datasets. The canonicalization of RDF graphs is known to be as difficult as the graph isomorphism problem, so depending on the input, there will inevitably be cases that require extremely long calculation times.

spouse's name while ensuring her digital signature remains valid, and thus create the verifiable RDF graph shown in Figure 10.

But suppose that a snooper who sees this graph knows only that Alice's spouse is either Bob or Charlie (but not which). This snooper could insert the names Bob and Charlie, first one and then the other, into the hidden part marked by *** and run the canonicalization process again, yielding two graphs with different canonicalized labels, as in Figure 11. By comparing these graphs with the graph that Alice published, the snooper would be able to determine that Alice's spouse is Bob.

This assumes special circumstances, but it is a property that you need to be aware of if using RDF Dataset Canonicalization with verifiable credentials. There is a discussion on avoiding this sort of problem within W3C Verifiable Credentials Data Integrity[*20], a specification for protecting the security and privacy of verifiable credentials.

2.8 Conclusion

I have provided an overview of RDF Dataset Canonicalization, now a W3C recommendation, looking at the specification itself and the standardization effort, which I was involved in. RDF Dataset Canonicalization makes it easier to calculate differences in RDF graphs, check for graph updates, calculate hashes, and generate digital signatures. This can streamline data management and make it possible to imbue RDF graphs with unforgeability and authenticity. As a user of the specification myself, I utilize it in research and development on verifiable credentials and their applications. I hope that this article has piqued your interest in the topic.

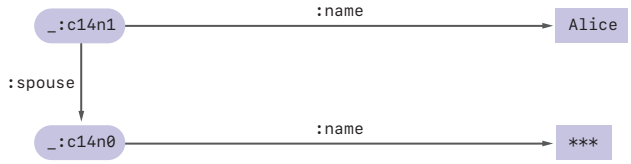


Figure10: RDF Graph with Bob's Name Hidden

Figure 11: Two Different Results

Dan Yamamoto
Senior Engineer, Office of Emergency Response and Clearinghouse for Security Information, Advanced Security Division, IIJ
Dr. Yamamoto has been in his current role since 2021. He is engaged in research on digital identity and information security.

*20 Manu Sporny, Dave Longley, Greg Bernstein, Dmitri Zagidulin, Sebastian Crane: Verifiable Credential Data Integrity 1.0. W3C Candidate Recommendation Draft, April 28, 2024 (https://www.w3.org/TR/2024/CRD-vc-data-integrity-20240428/).

IIJ's DRM Initiatives

3.1 Introduction

DRM stands for digital rights management and refers to technologies used to control copyrights by, for instance, imposing restrictions on the use and duplication of digital content. DRM applies to static content such as text and images as well as digital content in general, including music and games. In this article, I will discuss DRM as it relates to video delivery from the perspective of the IIJ Media Sphere Service, a video delivery platform that I work on at IIJ.

3.2 Overview of DRM

The digitization of music, video, and other content progressed rapidly in the 90s, and as the Internet became widespread in the latter half of that decade, the way such content was distributed changed dramatically. Local video rental stores turned into DVD rental stores, and these days we have online delivery services that let us easily watch what we want, when we want, wherever we want.

While this digitization of content brought considerable lifestyle benefits, it also made it simple to copy content without any degradation in quality, such that it was easy to envision piracy taking place. The distribution of illegal content is detrimental not only to the content rights holders but also to the companies involved in production, sales, delivery, and the various other aspects of the content business. Such behavior, if allowed to run rampant, threatens to send the industry itself into decline and, ultimately, discourage the production of engaging content. The end result of this is that we the consumers would no longer be able to enjoy such content.

It was against this backdrop that DRM technology was conceived as a means of protecting not only the content itself but also the development and advancement of the industry as a whole. The hope is that DRM will ensure the appropriate use of content and impose controls over acts such as the copying of that content.

It is crucial to note that DRM technology is not perfect. By its nature, it imposes restrictions on the content consumer's playback environment. There are inevitably cases in which even well-intentioned end users are unable to enjoy content depending on their playback environment. And several cases of DRM being misused have also been reported in the past.

Yet when it comes to Internet-based content delivery, DRM technology is now widespread and stable, and it is likely to grow in importance going forward.

3.3 The Evolution of DRM Services at IIJ

You may or may not be aware of this, but the DRM technology that is used to protect content is not all that new. While writing this article, I asked a long-time IIJ employee about this. He told me that he had already begun to perceive a need for DRM in the late 90s and had, accordingly, been gathering information at events such as the exhibitions run by Streaming Media[*1], a news media outlet that deals with online video delivery.

The first real service IIJ released was DRM for Flash Video in 2008[*2]. IIJ later added support for PlayReady, and in 2015 released a service that used the open-standard Marlin DRM system[*3].

The IIJ Media Sphere Service currently supports Apple's FairPlay Streaming, Google's Widevine, and Microsoft's PlayReady DRM systems. By supporting these three DRM systems, we believe we are at present able to cover a fairly comprehensive range of video playback environments.

*1 Streaming Media (https://www.streamingmedia.com/).

*2 IIJ, "IIJ Adds DRM Functionality to its Flash Video Delivery Solution" (https://www.iij.ad.jp/news/pressrelease/2008/pdf/FlashDRM.pdf, in Japanese).

*3 IIJ, "IIJ Begins Offering IIJ DRM Service/ExpressPlay® (https://www.iij.ad.jp/en/news/pressrelease/2015/0126.html).

By developing DRM functionality and providing it as a service within IIJ Media Sphere, IIJ has now made it easy for users to take advantage of DRM content protection without worrying about video player environments or content packaging. Not using an external DRM provider also makes it easier for IIJ to produce cost estimates. In terms of service sustainability as well, we believe that developing the system in-house and operating it within our own facilities will make it possible to deliver even greater reliability. This is because maintaining the underlying platform is key to guaranteeing an operating environment over the long term, and the same goes for the programs, including the source code. Going forward, we hope to continue to provide support and information about devices and create all sorts of value-added by implementing reporting, analysis, and other features.

Several employees at IIJ, including myself, have passed the Widevine certification program, enabling IIJ to become a Certified Widevine Implementation Partner.

3.4 DRM Features

The basic concept behind current DRM is to protect content by encrypting it and only allow it to be used in appropriate playback environments. DRM, as the name suggests, allows for a variety of specific rights management features to be used. These features are used by content providers and distributors in the form of policies that take the nature of their businesses into account.

Perhaps the most familiar such feature is video playback permissions, which make it possible to allow content playback only when your desired conditions are met. Another type of limitation that may also be familiar to you are those imposed on the number of devices that can simultaneously play back content.

You can also control content quality. For example, you can allow only audio playback, or manage the playback environment options—SD quality (480p), HD quality (1080p), and UHD quality (4k and above).

The same goes for HDCP. You may have seen the term HDCP in recent years when purchasing a device like a TV or smartphone, or in content delivery service literature. HDCP is a DRM component technology. It stands for High-Bandwidth Digital Content Protection and is a type of encryption technology developed by Intel in 2000 to protect copyright by preventing unauthorized copying. HDCP is used to encrypt digital interfaces like HDMI when transmitting video. If HDCP is not supported on both the video output device and the input device, content may not be playable or image quality may be limited. A simple use case, for example, is when you want to impose restrictions on output to analog devices. The current widespread version, HDCP 2.2, was released over a decade ago, so there is probably not too much to worry about in terms of getting it to work, but if you are trying to transmit video over something like HDMI and are having problems playing content, we recommend that you check for HDCP support on all of your devices.

Finally, let's look at device security levels. Widevine, for instance, specifies three security levels for devices: L1, L2, and L3. L1 is Widevine's most secure device level, and applies to devices that can decrypt and play videos within Trusted Execution Environment (TEE) hardware. L3 devices are those that do not have a TEE and thus perform decryption and video playback in software only. Based on these levels, it is possible to control playback so that only L1 devices are allowed to play back high-definition content, while L3 devices can only play back in low quality.

For Android smartphones, you can use an app called DRM Info to check the device's security level, so give it a try if you're interested.

3.5 How DRM Works

Below, I explain how DRM works, dealing first with the encryption process and then the decryption process.

3.5.1 Content Encryption

As I said, the basic concept behind DRM is to encrypt content. So, in general, during the video packaging process, the packaging system interacts with the DRM provider's key server when performing encryption. But there is a whole range of DRM systems out there, and thus a whole range of methods are in use. So a standard called CPIX (Content Protection Information Exchange), developed by the DASH Industry Forum (DASH-IF), has become widespread in the industry as a way of preventing packagers from having to deal with all these DRM systems separately. CPIX was originally developed for MPEG-DASH but now also supports HLS. The advantages of using CPIX are as follows.

1. Interoperability
   CPIX enables interoperability between different DRM systems. This allows content providers and delivery platforms to use multiple DRM providers, making it easy to deliver content across a range of devices and platforms.
2. Simplified workflow
   By adhering to CPIX, content providers and delivery platforms can use a common key and encryption format across multiple DRM systems. This makes key management and encryption procedures much simpler, resulting in more efficient workflows.
3. Enhanced security
   CPIX also offers security benefits. The use of a common encryption format and key mapping makes it easier to apply a consistent security policy to prevent content breaches and unauthorized copying.
4. Open standard
   CPIX is an open standard and has been adopted industry-wide. This helps prevent vendor lock-in*4.

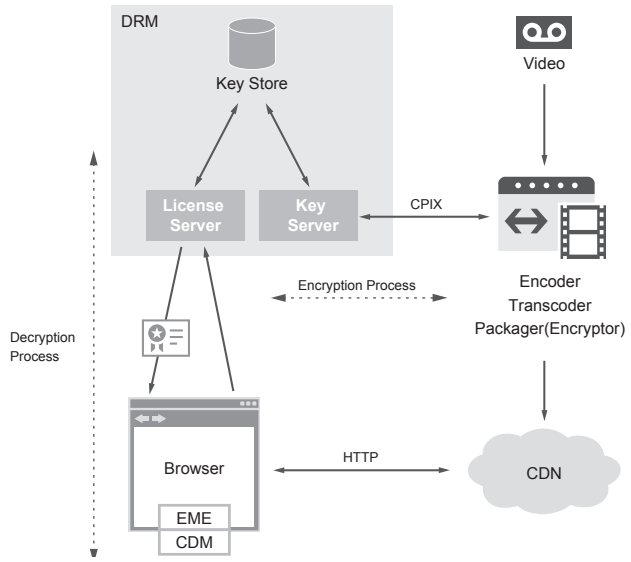


Figure 1: The Encryption and Decryption Processes

*4 DASH Industry Forum, DASH-IF Implementation Guidelines: Content Protection Information Exchange Format (https://dashif.org/docs/CPIX2.2/Cpix.pdf).

Let's look at an example.

```
  '
    <cpix:CPIX id="cpixsample" xmlns:cpix="urn:dashif:org:cpix"
xmlns:pskc="urn:ietf:params:xml:ns:keyprov:pskc">
      <cpix:ContentKeyList>
        <cpix:ContentKey kid="f269e534-c4f1-4721-9d62-26dc7ed241bd"
explicitIV="8mnlNMTx...">
        <cpix:Data>
          <pskc:Secret>
           <pskc:PlainValue>vfMB2...</pskc:PlainValue>
          </pskc:Secret>
        </cpix:Data>
      </cpix:ContentKey>
      </cpix:ContentKeyList>
      <cpix:DRMSystemList>
        <cpix:DRMSystem kid="f269e534-c4f1-4721-9d62-26dc7ed241bd"
systemId="edef8ba9-79d6-4ace-a3c8-27dcd51d21ed">
          <cpix:PSSH>AAAAP3Bzc...</cpix:PSSH>
        </cpix:DRMSystem>
        <cpix:DRMSystem kid="f269e534-c4f1-4721-9d62-26dc7ed241bd"
systemId="9a04f079-9840-4286-ab92-e65be0885f95">
          <cpix:PSSH>AAADBnBzc...</cpix:PSSH>

<cpix:ContentProtectionData>PG1zcHI6cHJvIHhtbG...</cpix:ContentProt
ectionData>
        </cpix:DRMSystem>
        <cpix:DRMSystem kid="f269e534-c4f1-4721-9d62-26dc7ed241bd"
systemId="94ce86fb-07ff-4f43-adb8-93d2fa968ca2">
          <cpix:URIExtXKey>c2tkOi...</cpix:URIExtXKey>
        </cpix:DRMSystem>
      </cpix:DRMSystemList>
    </cpix:CPIX>
  `
```

CPIX is written in XML. I will now go over a number of key points using the example above.

The ContentKey element represents the information required for content encryption. This is based on an extends RFC 6030 Portable Symmetric Key Container (PSKC)[*5].

The DRMSystem elements represent information specific to each DRM system. The systemId strings are IDs that identify each DRM system and are determined by DASH-IF[*6].

Intheexampleabove,edef8ba9-79d6-4ace-a3c8-27dcd51d21ed referstoWidevine,9a04f079-9840-4286-ab92-e65be0885f95 refers to Microsoft PlayReady, and 94ce86fb-07ff-4f43-adb8-93d2fa968ca2 refers to Apple FairPlay.

The PSSH element represents data used in the MP4 PSSH (Protection System Specific Header) box, a type of video container. This is a type of MP4 box used to store information related to digital content encryption and DRM systems. The PSSH box contains the encryption key, the encryption method, and other information about the DRM system.

The URIExtXKey element, as you can probably tell from the fact that the relevant DRM system is used by Apple

*5 RFC Editor, RFC 6030 Portable Symmetric Key Container (PSKC), October 2010 (https://www.rfc-editor.org/info/rfc6030).
*6 DASH Industry Forum, Content Protection (https://dashif.org/identifiers/content_protection/).

FairPlay, is an item that affects the EXT-X-KEY used in HLS playlists.

This information is used during the content packaging process to encrypt the content.

The SPEKE (Secure Packager and Encoder Key Exchange) standard, which extends CPIX and is available on AWS, is also very widespread. This is also an open specification, so anyone can figure out how it works[*7].

3.5.2 Content Decryption

So at this point, our content has been encrypted. We are probably seldom aware of the existence of DRM when casually watching videos in a browser or the like. Yet, some complex processing needs to take place for us to play encrypted content. This section gives a simple explanation of what happens.

A W3C specification called EME (Encrypted Media Extensions), which provides a communication channel

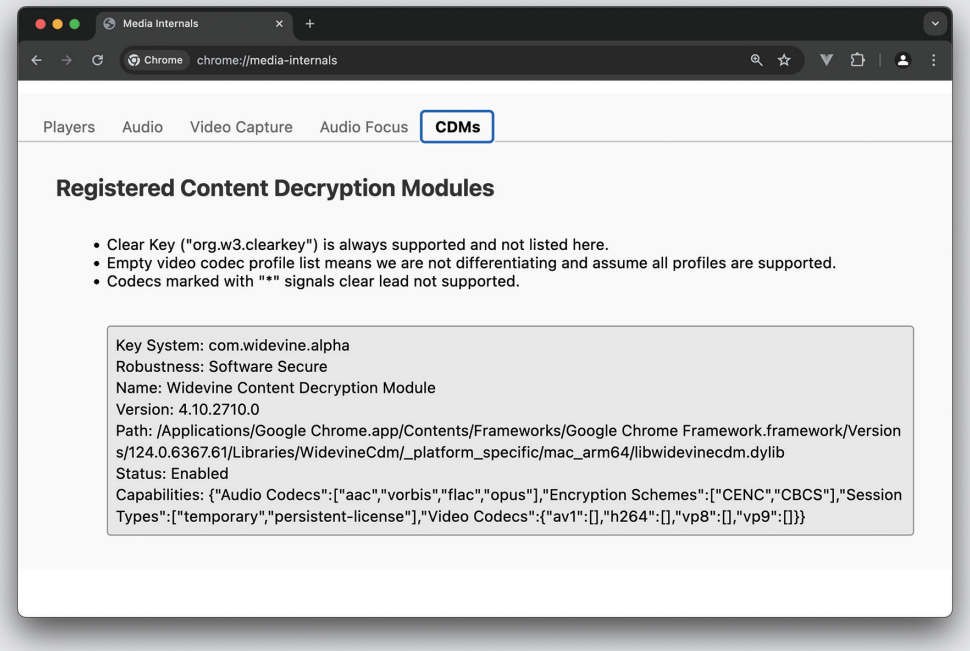


Figure 2: Chrome's media-internals

*7 AWS, SPEKE API specification (https://docs.aws.amazon.com/speke/latest/documentation/speke-api-specification.html).

between web browsers and DRM software, and the CDMs (Content Decryption Modules) provided by DRM vendors in accord with that specification, are key to decrypting the content.

CDMs are closed source, and Widevine, for example, provides a CDM in the form of a plugin for Google Chrome as well as for Firefox (Figures 2 and 3). Apple FairPlay's CDM is only available for Apple products, such as Safari, so it cannot be used on Google Chrome and other browsers.

When the EME specification was being developed, CDMs were the subject of strong opposition from supporters of an open web, but in the end, the specification was formulated in such a way that the CDMs would only be responsible for confidentiality and integrity for the purposes of decryption, while the video players and other such applications would be in charge of the data communications. If EME and CDM had not been formulated in this way, DRM providers would not have been able to take a consistent approach, and content delivery services may all have had to use their own specific applications.

In Google Chrome, you can view detailed information on the CDMs by typing "chrome://media-internals/" into the address bar, so take a look if you're interested.

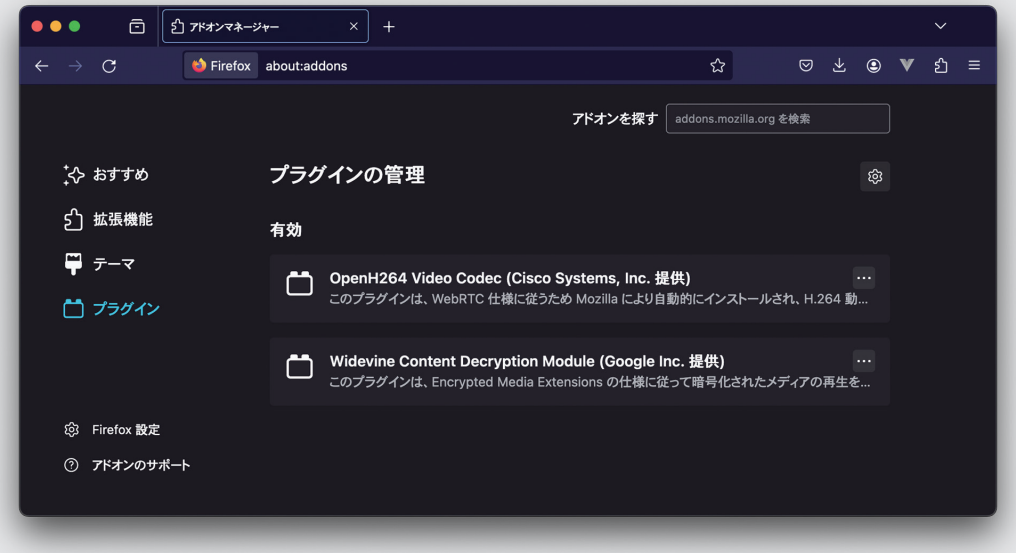
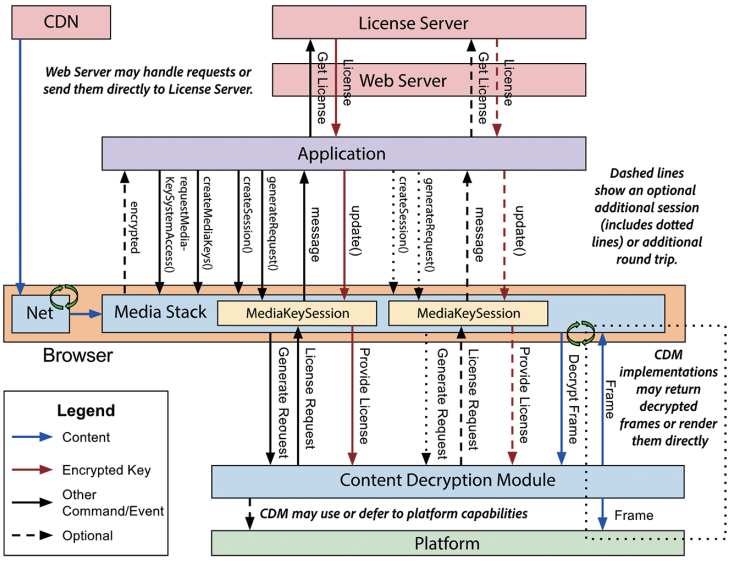


Figure 3: Firefox plugins

Figure 4 shows the decryption flow when using EME and CDM. Video players are only able to play content after communicating with a CDM using EME. This involves the following steps.

1. The browser fires an encrypted event, which tells the video player that the content is encrypted, and the video player calls the requestMediaKeySystemAccess() method to use a specific CDM.
2. The createMediaKeys() method initializes the keys.
3. The createSession() method creates a session to control key expiration time.
4. The generateRequest() method tells the CDM to generate a license request. The license is the most important piece of data, containing the decryption key and information needed for using DRM features.
5. The CDM generates and returns a license request. The video player learns about the license request via a message event.
6. The video player sends the license request to the license server.
7. The video player receives the license returned by the license server.
8. The returned license is passed to the CDM using the update() method.

The CDM assesses the returned license based on the DRM policy and decrypts the content, and the video is thus played.



Source: W3C, Encrypted Media Extensions (https://www.w3.org/TR/encrypted-media/).

Figure 4: Decryption Flow with EME and CDM

3.6 Conclusion

In this article, I have discussed DRM as it relates to video delivery, with a look at the development of the IIJ Media Sphere Service. These days, no small number of people enjoy video content via content delivery services. I have endeavored to explain how DRM technology is actually used behind the scenes in this space and roughly how DRM works. While I have been somewhat constrained by space limitations, I thank you for reading, and I hope you found this informative.

Mitsuo Kuroishi
Deputy Manager, Delivery Development Section, Content Delivery Services Department, Network Division, IIJ
Mr. Kuroishi has worked on the development of a range of services since joining IIJ in 2002. His favorite saying is “Code speaks louder than words,” and he looks forward to Emacs updates.

August 2024 Vol.63

Internet Initiative Japan

About Internet Initiative Japan Inc. (IIJ)

IIJ was established in 1992, mainly by a group of engineers who had been involved in research and development activities related to the Internet, under the concept of promoting the widespread use of the Internet in Japan.
IIJ currently operates one of the largest Internet backbones in Japan, manages Internet infrastructures, and provides comprehensive high-quality system environments (including Internet access, systems integration, and outsourcing services, etc.) to high-end business users including the government and other public offices and financial institutions.
In addition, IIJ actively shares knowledge accumulated through service development and Internet backbone operation, and is making efforts to expand the Internet used as a social infrastructure.

IIJ-MKTG020-0061

Internet Initiative Japan Inc.

Address: Iidabashi Grand Bloom, 2-10-2 Fujimi, Chiyoda-ku,
Tokyo 102-0071, Japan
Email: info@iij.ad.jp URL: https://www.iij.ad.jp/en/