Security and Privacy Considerations of Metadata

Radhika Ravindranath
New Jersey, USA
(Email: radhika.ravindranat@gmail.com)

--************************----------------------------------

Abstract:

Metadata or meta information, also known as “data that provides information about other data”[1] provides information to help users and programs to classify, analyze, label, categorize and search through data assets. While metadata is invaluable in processing information and maintaining compliance and regulatory standards, it comes with its own set of security and privacy challenges. Metadata can be a treasure trove of personal information. This whitepaper elucidates the nature of metadata, types and risks. Metadata can pose several risks to an organization, including reputational damage, regulatory fines and financial losses. This paper aims to delve into key security and privacy risks associated with the collection, processing and storage of metadata. Ways to mitigate these risks to protect the organization and its users are discussed. In conclusion, metadata is essential to the way organizations and platforms store, organize and maintain data. The proper management of metadata can enable organizations to not only reduce the security and privacy risks associated with metadata, but to enhance their overall security and privacy posture.

Keywords - metadata, security, privacy, data governance, data protection, metadata management

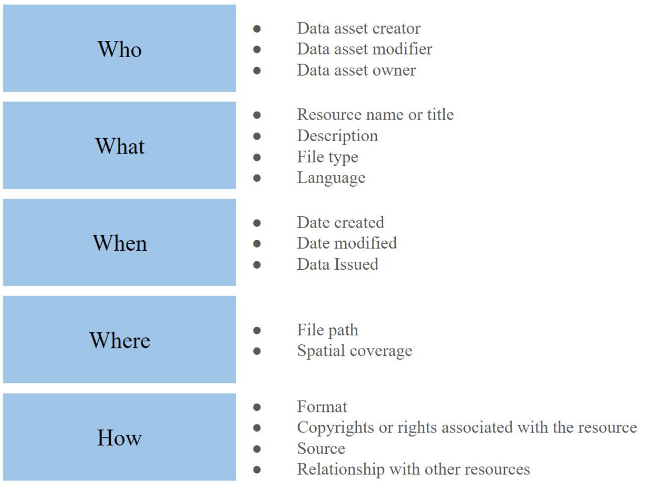
--************************----------------------------------

I. INTRODUCTION

Metadata is data that provides information about data[1], and serves as a critical component in information systems management.In general, metadata is used to provide context about the structure of digital assets and their classification.

User specific metadata, which is predominantly referenced in this paper, is directly associated with content generated by a user, or associated with a user when combined with other data sources. It can be used to glean the ‘who’, ‘what’, ‘how’, ‘when’ and ‘where’ of a data element.

For example, in the case of making a phone call, the metadata generated may include caller and receiver numbers(who), phone(what) MAC addresses(how), time and duration(when) of the call and the location(where) of the caller and receiver[2].



Who	Data asset creator Data asset modifier Data asset owner
What	Resource name or title Description File type Language
When	Date created Date modified Data Issued
Where	File path Spatial coverage
How	Format Copyrights or rights associated with the resource Source Relationship with other resources

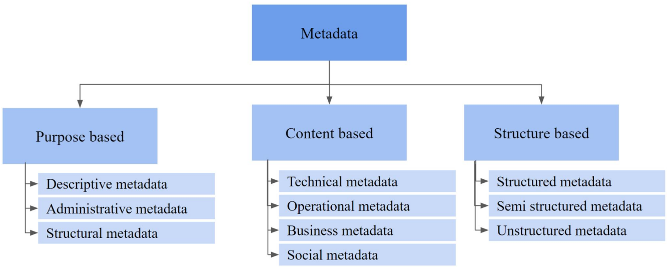
While the generation of metadata may be done through lawful purposes, there is an enduring debate as to how metadata must be treated in order to protect users from security and privacy risks.

The General Data Protection Regulation (GDPR) defines personal data as any information that relates to an identified or identifiable natural person. While a single piece of metadata from a single transmission may not be considered identifiable personal data, amassing a large amount of metadata could be used to identify a person through a process known as reidentification. Metadata can be used in processes like "social network analysis" or "contact chaining" to draft the network of an individual upto one or two degrees of separation as provided in the example of "Connecting the Dots: Tracking Two Terrorist Suspects"[3] .Granularity of metadata can also be privacy and security impacting. For example, a phone call made from a device in a sparsely populated area could be used to hone in on a single individual in that area if combined with other pieces of information, such as a list of individuals in that area. If metadata can be used to identify or track an individual, it is considered personal data under the GDPR[4].

I. UNDERSTANDING METADATA

The first step to analyzing the risks of metadata is to map out the information conveyed by it.

By analyzing different sources, this whitepaper provides some ways in which metadata can be bucketed. Metadata can be classified based on purpose, content and structure. Purpose based classification includes descriptive metadata, administrative metadata and structural metadata[5]. Content based metadata can include technical metadata, operational metadata[6], business metadata[7] and social metadata[8]. Structure based classification includes structured metadata, semi structured metadata and unstructured meta data[9].



Purpose based metadata can be further sub categorized into a number of buckets. Descriptive metadata describes the characteristics of a data asset, such as title, author, date and subject. Administrative metadata provides information on who has access to the resource and when it was created, deleted or modified. Structural metadata may show the hierarchy of a document or nesting in a webpage. Technical metadata can provide information on a data asset such as size of a file, retention period on a bucket, etc.

Content based classifications describe the details of a resource, the entities the resource is dependent on - like affiliated people, places and organizations - and the relationship to other entities like which databases it may be connected to, which assets data is shared with, etc.

Structure-based metadata includes information of how the metadata is structured. This may be useful in the storage and processing of the metadata. For example, structured metadata may be organized in the form of JSON or XML, Unstructured metadata includes free text descriptions, and semi structured metadata includes key value pairs.

Ultimately, the classification of metadata can vary based on its applications. Some examples of industry standards which proscribe metadata definitions are EXIF (Exchangeable File Format), which is a standard for storing metadata in image files and includes technical data like camera settings, date and location. Extensible Metadata Platform (XMP) is a standard by Adobe for storing metadata about digital files such as licensing, copyright and usage rights.

Schema.org and NISO are popular collaborative and industry standards which can be used to structure and standardize metadata for the proper organization of data assets.

II. SECURITY RISKS OF METADATA

Once the type of metadata, its contents and the purpose for which it is collected are understood, the security risks associated with it become more apparent. Below are some of the primary security risks associated with metadata;

A. Data Exposure

Metadata can reveal personally identifiable or sensitive information. For example, descriptive filenames like “confidential_honeypotIPs.docx” can reveal content or privilege of a document. Timestamps can reveal valuable information about activity patterns, which can be used in social engineering attacks. Geolocation data embedded in photos and videos can be used to compromise security and privacy. Additionally some metadata formats can be used for the covert storage of “hidden data[10]”, which can contain sensitive information used for tracking and surveillance in seemingly innocuous data assets like videos or images[11].

B. Unauthorized Access

Attackers can gather valuable information to identify targets. It can also be used to perform vulnerability assessments to determine the structure and organization of a system, and to locate specific files and databases containing sensitive information for targeted attacks. Exploiting vulnerabilities in Cloud Instance metadata APIs[12] can also allow attackers to gain unauthorized access to sensitive information or credentials.

C. Data Breaches

Mismanaged metadata can pose several security risks, exposing sensitive data such as the location of financial records and intellectual property to potential extraction and exfiltration. For instance, the AT&T data breaches of 2022 and 2023 compromised vast amounts of metadata, including caller information, making its users vulnerable to exploitation by nation-state actors, criminal organizations, and rival businesses.[13]

D. Denial of Service (DoS) attacks

Attack techniques like metadata based flooding can disrupt normal operations leading to denial of service attacks. Malicious metadata injection into a system leading to system instability or security breaches is another mode of DoS used by attackers[14].

III. PRIVACY RISKS OF METADATA

Even when the content of communications are encrypted, metadata can be collected and analyzed to breach the privacy of individuals in an organization. This information can be used to create detailed profiles of a user's online presence, enabling surveillance, targeted advertising and discrimination[15].

A. Disclosure of Personal Information

Communication patterns like frequency, duration and timing of communications can reveal personal habits and strength of relationships.

Location history, search queries and social media interactions can reveal personal interests, opinions and beliefs of an individual without their consent[16].

B. Profiling and tracking

Cookie data, while independently not personally identifiable, can be combined with other metadata to build detailed profiles about an individual, their browsing habits, social interactions, demographics and behaviors[17]. Metadata containing location information may be used for profiling and tracking purposes.

C. Surveillance

Metadata can allow companies and governments to engage in surveillance activities. For example, PRISM, a US government surveillance program in 2013 collected vast amounts of content and metadata[18]including email traffic, phone calls and email traffic. Companies use first party and third party cookies to track user patterns across websites for targeted advertising[19].

D. Discrimination

Metadata may also be used to discriminate against individuals and their access to certain services. For example, IP addresses can be used to determine the nationality of an individual, and subsequent services like the cost of internet, or access to wifi can be curbed. Metadata can further perpetuate algorithmic bias when used to train machine learning models[20].

IV. MITIGATION STRATEGIES

A comprehensive approach is necessary to mitigate the security and privacy risks associated with metadata.

A. Understand the metadata

Recognizing the different types of metadata and identifying the potential risks associated with each type is a crucial first step in mitigating metadata related security and privacy risks[21]. Determining the sources of metadata in system logs, user generated content, and external data sources and managing them accordingly is essential in effective security and privacy practices.

B. Treat Metadata as a Data Asset:

Manage metadata based on sensitivity levels such as “public”, “internal”, “confidential”, “restricted”, “mission critical” and apply the appropriate security controls like data masking on that data based on classification levels. Periodically assess metadata for potential vulnerabilities like exposure to unauthorized access and privacy violations. Develop organizational policies and technical controls to protect metadata.

C. Access control

Ensure that access to metadata is restricted to authorized personnel. Grant users only the limited amount of privileges needed to do their jobs.

D. Scrubbing and Anonymization

Develop clear policies for metadata access permissions and remove sensitive information from metadata before storing and sharing. For example, sensitive field names which are not required for analysis must be excluded in logging systems, anonymization or pseudonymization of free text fields used for analysis to prevent data breaches.

E. Metadata management tools

Leverage data cataloging and metadata management tools like Collibra, Informatica and Atlan to discover, classify and govern data assets. Logging systems like Splunk and Chronicle allow administrators to identify patterns in the logs for scrubbing sensitive field information. Leveraging these tools will allow organizations to create sensitivity labels, automate classification, access controls and tagging tasks and improve efficiency by reducing manual effort. They can also allow administrators to gain valuable insight into data usage patterns and potential anomalies[22].

V. CONCLUSIONS

In conclusion, metadata is a powerful tool which can be used to organize, manage and understand digital information. While metadata by itself may not reveal the core information in a message, it can reveal sensitive information which can be used in credential theft and data breaches. Additionally, metadata can be used for denial of service attacks, discrimination and surveillance. To mitigate these risks, it is important for organizations to implement robust security measures such as data classification, access controls, encryption and regular audits. By treating metadata as a valuable asset and taking proactive steps to protect it, organizations can ensure the security and privacy of their digital information.

REFERENCES

[1] Merriam-Webster [Online]. Available: https://www.merriam-webster.com/dictionary/metadata

[2] Office of the Privacy Commissioner of Canada (2014). Metadata and Privacy - A Technical and Legal Overview [Online]. Available: https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2014/md_201410/.

[3] Office of the Secretary of Defense ([n.d.]). Connecting the Dots -- Social Network Analysis of 9-11 Terror Network [Online]. Available: http://orgnet.com/tnet.html.

[4] Metadata.io (2016). GDPR Policy [Online]. Available: https://metadata.io/gdpr-policy/

[5] Chia, A. (2023, June 22). Types of Metadata [Online]. Splunk. Available: https://www.splunk.com/en_us/blog/learn/metadata-types.html#:~:text=Descriptive%20metadata%20describes%20a%20resource's,Date%20of%20modification

[6] Bhansali, N. (Ed.). (2013). Data Governance: Creating Value from Information Assets. CRC Press.

[7] Inmon, W. H., O'Neil, B., & Fryman, L. (2010). Data Governance: Creating Value from Information Assets. Morgan Kaufmann.

[8] Klas, W. (2008). Metadata for Semantic and Social Applications. In Proceedings of the International Conference on Dublin Core and Metadata Applications, Berlin, 22 - 26 September 2008, DC 2008: Berlin, Germany.

[9] Inmon, W. H., O'Neil, B., & Fryman, L. (2010). Business Metadata: Capturing Enterprise Knowledge. Morgan Kaufmann.

[10] Hassan, N. A., & Hijazi, R. (2016). Data Hiding Techniques in Windows OS: A Practical Approach to Investigation and Defense.

[11] Smith, J. (2023). The risks of metadata exposure. Office of the Privacy Commissioner of Canada. Available: https://www.priv.gc.ca/en/privacy-topics/technology/02_05_d_30/

[12] MITRE ATT&CK [Online]. Unsecured Credentials: Cloud Instance Metadata API, Sub-technique T1552.005. Available: https://attack.mitre.org/techniques/T1552/005/

[13] CSO Online. (2017, June 22). AT&T's data breach isn't trivial, especially to spy agencies [Online]. Available: https://www.csoonline.com/article/2516887/atts-data-breach-isnt-trivial-especially-to-spy-agencies.html

[14] MITRE ATT&CK [Online]. Endpoint Denial of Service: Application or System Exploitation. Available: https://attack.mitre.org/techniques/T1499/.

[15] U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics [Online]. (2009). Forum Guide to Metadata – Chapter 3. Using Metadata – Data Profiling. Available: https://nces.ed.gov/pubs2009/metadata/ch3_profiling.asp

[16] Freedom House [Online]. (n.d.). Metadata - 102. Available: https://freedom.press/digisec/blog/metadata-102/

[17] Kaspersky Lab [Online]. What Are Internet Cookies and What Do They Do? Available: https://usa.kaspersky.com/resource-center/definitions/cookies

[18] "FBI, CIA Use Backdoor Searches To Warrentlessly Spy On Americans' Communications". TechDirt. June 30, 2014. Archived from the original on February 19, 2015. Retrieved February 19, 2015.

[19] Harvard Online [Online]. (2024, February 2). How Google's 2024 Third-Party Tracking Cookie Removal Impacts Users. Available: https://www.harvardonline.harvard.edu/blog/how-googles-third-party-tracking-cookie-removal-impacts-users-2024

[20] Civil Rights Org [Online]. Address Data-Driven Discrimination to Protect Civil Rights. Available: https://civilrights.org/resource/address-data-driven-discrimination-protect-civil-rights/

[21] Riley, J. (2017). Understanding Metadata: What is Metadata, and What is it For?: A Primer [Online]. National Information Standards Organization (NISO). Available: https://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf

[22] Simons, A., & vom Brocke, J. (Eds.). (2013). Enterprise Content Management in Information Systems Research: Foundations, Methods and Cases. Springer Berlin Heidelberg.

Mail your Manuscript to
editorijctjournal@gmail.com
editor@ijctjournal.org