Blinding Identity Taxonomy

Version:	1.0
Document Date:	2020-06-15
Editors:	Paul Knowles, John Wunderlich, Ken Klingenstein
Contributors:	See Section 6.2 Participants
Produced by:	ISI Work Group Participant Roster Information Sharing Interoperability Work Group https://kantarainitiative.org/groups/isi-work-group/

Status:
This document is a Group-Editors' DraftReport produced by the Information Sharing Interoperability Work Group. See the Kantara Initiative Operating Procedures for more information on Kantara Reports, Recommendations and Specifications.

Abstract:
The BIT is a taxonomy of data fields to be blinded for the purpose of removing identity data from a dataset.

Suggested Citation:
Blinding Identity Taxonomy 1.0. Kantara Initiative Information Sharing Interoperability Work Group. 2020-06-15. Kantara Initiative Report.

NOTICE

IPR Option: Non-Assertion Covenant

DEAR READER

Thank you for downloading this publication prepared by the international community of experts that comprise the Kantara Initiative. Kantara is a global non-profit 'commons' dedicated to improving trustworthy use of digital identity and personal data through innovation, standardization and good practice. 'Nurture, Develop, Operate' - that's what Kantara does.

Kantara is known around the world for incubating innovative concepts, operating Trust Frameworks to assure digital identity and privacy service providers and developing community-led best practice and specifications. Its efforts are acknowledged by OECD ITAC, UNCITRAL, ISO SC27, other consortia and governments around the world

Every publication, in every domain, is capable of improvement. Kantara welcomes and values your contribution through membership, sponsorship and active participation in all our endeavors including in the working group that produced this publication. With your contribution, Kantara can reflect its value back to you and your organization while continuing to consolidate an inclusive, equitable digital economy offering value and benefit to all.

Contents

Tables

1 EXECUTIVE SUMMARY

Identifiers and attributes are the payloads that are exchanged in the identity landscape. Identifiers are assumed to have some correspondence with the entity they refer to. Attributes are generally seen to be characteristics possessed by multiple entities. When entities have a business requirement to process information they may or may not require identifiers or identifying attributes. From a privacy perspective, processing or sharing fewer identifiers or identify attributes reduces risk. Unauthorized access or use to a dataset exposes identifiers and hence identities. Correlation attacks attempt to use multiple attributes of an end-entity to identify an individual. Meta-data about a dataset can be mined to infer personal information.

The purpose of BIT is to provide a defensive tool against re-identification attacks against a dataset AND to make Datasets more useful for a range of purposes. The Blinding Identity Taxonomy (BIT) outlined in this Kantara Report is therefore intended to be a practical tool for practitioners whose organization have custody or control of a dataset that contains identifiable information about entities. These entities may be, but are not necessarily, natural persons. The purpose of the tool is to provide the ability of an organizations that wants to store, use or share datasets with a reasonable claim that the dataset does not contain identifying information about entities. A reasonable claim may depend on the use of additional controls in addition to the application of the BIT, such as contracts or security controls. These additional controls are out of scope of this report.

The BIT is a list of field and categories of fields that may be contained in a dataset. Where a real-world instance of a dataset contains these fields, they will be blinded by being encrypted or excluded from the ‘blinded’ dataset in such a way that the resulting dataset no longer contains readable data in any of the defined fields. See BIT Fields & Notes below for details on the BIT.

The Kantara Initiative is publishing this report as a service to the identity community to provide another tool for identity practitioners looking for practicable methods to reduce risk and meet business goals. We hope that practitioners that can use this taxonomy are able to share their experiences with the community to be able to identify common profiles and schemas that will facilitate adoption of the Blinding Identity Taxonomy. Version 1.0 of the report includes a high-level view of profiles and schemas that practitioners can use for their particular use cases.

2 INTRODUCTION

The Blinding Identity Taxonomy (BIT) was first created in 2018 with the aim of providing a common reference or practice to enable the protection of identities. For the purposes of this report identity refers to the attributes of a natural person or organization or a device with signing capabilities that make that entity uniquely identifiable[1]. BIT can be used to flag a list of elements which require cryptographic encoding to reduce the risk of identifying a data principal. When those elements have been removed or encrypted, the dataset is 'blinded'. For the purposes of this report, a dataset may be said to be successfully 'blinded' when an adversary with access to the dataset cannot identify a significant number of the data principals contained in the dataset. This recognizes that no blinding effort is risk free and allows blinding organizations to determine the level of risk of unblinding that is appropriate for their contexts.

This report is being published by the Kantara Initiative as a resource to the identity and information sharing community to assist policy makers and technologists to make decisions about where and how to apply blinding techniques to Datasets with identity attributes. The authors hope that real world use cases, profiles and schemas will be contributed to future versions.

[1] ISO 20889 refers to this as a data principal

3 BLINDING IDENTITY TAXONOMY

3.1 SCOPE

This report provides a description of the Blinding Identity Taxonomy (BIT), to be used for the purpose of blinding datasets by removing or encrypting fields containing identifying information about data principals. The report assumes that an entity has custody or control of a dataset that contains identifying information about entities. The goal of entities that choose to use the BIT will be to process that dataset, or a copy of it, with the ability to make a reasonable claim that the resulting 'blinded' dataset does not contain identifying information about data principals. A 'reasonable' claim may depend on other controls, such as administrative, physical or technical controls that are out of scope of this report.

3.2 DEFINITIONS

For the purpose of this report the following definitions apply. Where possible existing standard definitions are referenced[2].

Term	Definition
Aggregated data	Data representing a group of data principals, such as a collection of statistical properties of that group. Source: ISO/IEC 20889:2018, 3.2
Attribute	Inherent characteristic. Source: ISO9241-302:2008, 3.4.2
Blinded	A dataset from which identifiers and quasi-identifiers have been removed or altered to reduce the risk that records can be associated with the entity referred to by that record in a given operational context. Note: Determining the amount of risk reduction that is acceptable in a given context is out of scope of this report.
Blinding	A technique that results in a blinded dataset.
Data Principal	The entity to which data relates. Source: ISO/IEC 20889:2018, 3.4
Dataset	Collection of data. Source: ISO/IEC 20889:2018, 3.5
Direct Identifier	An attribute that alone enables the unique identification of a data principal within a specific operational context
Equivalence class	A set of records in a dataset that have the same values for a specified subset of attributes. Source: ISO/IEC 20889:2018, 3.11
Identifier	A set of attributes in a dataset that enable unique identification of a data principal within a specific operational context. Source: ISO/IEC 20889:2018, 3.13
Identifying attribute	An attribute in a dataset that can contribute to uniquely identifying a data principal with a specific operational context. Source: ISO/IEC 20889:2018, 3.14
Indirect identifier	An attribute that, together with other attributes that can be in a dataset or external to it, enable unique identification of a data principal within a specific operational context. Source: ISO/IEC 20889:2018 3.16
Pseudonym	A unique identifier created for a data principal to replace the commonly used identifier or identifiers for that data principal. Source: ISO/IEC 20889:2018 3.26
Pseudonymization	A de-identification technique that replaces an identifier or identifiers for a data principal with a pseudonym to order to hide the identity of that data principal. Source: ISO/IEC 20889:2018 3.27
Quasi-identifier	An attribute in a dataset that, when considered in conjunction with other attributes in the dataset, singles out a data principal. Source: ISO/IEC 20889:2018 3.28
Record	A set of attributes concerning a single data principal. Source: ISO/IEC 20889:2018 3.30

[2] For ISO terms and definitions, check out their online browsing platform at https://www.iso.org/obp/ui and select "Terms and Definitions."

Term	Definition
Sensitive attribute[3]	An attribute in a dataset that, depending on the application context, merits specific, high-level protection against re-identification attacks enabling disclosure of its values, its existence, or association with any of the data principals. Source: ISO/IEC 20889:2018 3.34
Single out	To isolate records belonging to a data principal in the dataset by observing a set of characteristics known to uniquely identity this data principal. Source: ISO/IEC 20889:2018 3.35
Unique identifier	An attribute in a dataset that alone singles out a data principal in the dataset. Source: ISO/IEC 20889:2018 3.39

[3] This should be distinguished from sensitive in a regulatory sense relating to the psychological or other impacts on the data principal. Sensitivity here relates to re-identification risk.

4 PROFILES AND SCHEMAS

The Blinding Identity Taxonomy (BIT) can be used to select the fields that should be encrypted except where those fields are in use and unblinding is required for the use. In this context the BIT can determine what should be encrypted at rest and in motion. Another use of the BIT is to use profiles and schemas to enable identity safe uses of data.
By using profiles and schemas practitioners may be able to automate the application of the BIT and produce multiple blinded datasets relatively easily. A typical use case that might lead to a profile is this. A medical researcher wants to do a chart review for a research study. As part of their ethical review process they determine that they can use blinded data because they are only interested in clinical data, not identifying data. The researcher reviews the data dictionary or fields available to them in charts (the field definitions, not the contents) resulting in a description of the fields that they would like to 'pull' for their chart review. They then review the fields against the BIT and determine if they want to use a field that should be blinded. They can then choose a number of options for blinding that field, including format preserving encryption or by combining a set of attributes into a single attribute (replacing full date of birth with age at start of study period, for example). Once the blinding process has been identified it can be applied to the original data set to produce the blinded data set for review.

Should the process prove useful, this blinding process can then be generalized and captured in a profile that defines the fields that will be extracted and a description of the blinding process on a field by field basis. Where the extract is in a defined structured format like JSON or JSON-LD, a schema for that data export can also be attached.

4.1 PROFILES

Profiles are subsets of available data. A profile will typically be a list of fields, with a specification of each field type and its characteristics for inclusion in the destination dataset. For fields that are listed above as part of the taxonomy and that will be included in the destination dataset, the profile may recommend a type of encryption.
When data is collected or captured a profile will be used to select and or transform the data that is added to a dataset or database. In some contexts, profiles can be used as a form of data minimization to ensure that the only data that will be entered into a dataset or database is data that is necessary for the purpose for which the dataset or database was created. In all cases, profiles will enable more effective ETL[4] routines.
In a medical research or clinical assessment context for example, a profile sets out the elements of a patient's health record that are required for a research project or assessment. To the extent feasible in the context this may also converting free form text fields to defined content fields to better enable machine processing using schemas. An example would be the conversion of a text diagnosis to its corresponding ICD code[5].
Profiles will, if shared, enable blinded data interoperability between organizations.

4.2 SCHEMAS

A schema is a machine-readable data structure that defines the semantics of the data contained in the structure. A well-defined schema will contain a group or groups of related attributes that, when amalgamated, will provide a concise context which can be summarised and captured in the metadata block of the data structure.

[4] Export, Translate, Load
[5] At the time of writing ICD-10 is used by WHO member states for reporting, while ICD-11 has been released for preparatory purposes.

5 CONCLUSION

The Blinding Identity Taxonomy is a tool to enable implementers to safeguard their data from unauthorized or inappropriate uses. The use of BIT as a de-identification technique will enable implementers to provide stakeholder assurances about their datasets. We hope that this will prove to be useful and practical guidance for implementers.

6 APPENDICES

6.1 HOW THE BIT CAME TO BE

The Blinding Identity Taxonomy (BIT) was conceived on April 4th, 2018 on the second day of the 26th edition of the Internet Identity Workshop (IIW) at the Computer History Museum in Mountain View in a private conversation regarding the technical limitations of the EU's General Data Protection Regulation (GDPR) which was due to be enforced the following month on May 25th.

The first draft of listed elements was produced by Paul Knowles, Jan Lindquist (then OTT Analytics Specialist, Dativa) and Tom Weiss (CTO and Chief Data Scientist, Dativa) following blanket review from various members of the Identity and Big Data communities and was subsequently published as a Dativa blog post titled "The blinding identity taxonomy initiative" on September 6th, 2018[6] which was authored by Paul Knowles.

The intellectual property rights for the BIT were transferred to Kantara Initiative's Consent Information Sharing work group (CIS-WG) in December 2018 and re-contributed to the newly formed Information Sharing Interoperability work group (ISI-WG) by Paul Knowles[7] and Jan Lindquist[8] on January 22nd, 2020.

6.2 PARTICIPANTS

The following individuals participated in the creation of this document. For a complete list of participants see the Information Sharing Interoperability Work Group (WG-ISI) Participant Roster: https://kantarainitiative.org/confluence/display/WGISI/Participant+Roster

Individual	Organization
Iain Henderson	Individual Contributor
Jim Pasquale	digi.me
John Wunderlich	Individual Contributor
Kenneth Klingenstein	Internet2
Lisa LeVasseur	Individual Contributor
Mary Hodder	Individual Contributor
Paul Knowles	The Human Colossus Foundation
Salvatore D'Agostino	Individual contributor

Table 1 Participant List

6.3 REFERENCES AND USEFUL LINKS

Document/Reference	Short URL Link
ISO/IEC 20889 Privacy enhancing data de-identification terminology and classification of techniques	https://bit.ly/3cEk0T2
ISO/IEC 27000 Information technology – Security techniques – Information security management systems – Overview and vocabulary	https://bit.ly/2ZbArT4
ISO/IEC 29100 Information technology – Security techniques – Privacy framework	https://bit.ly/364mqb7
International Statistical Classification of Diseases and Related Health Problems	https://bit.ly/2X6BRLG

Table 2 References and Useful Links

[6] https://www.dativa.com/blogs/blinding-identity-taxonomy/
[7] https://kantarainitiative.org/confluence/display/WGISI/Re-contributions+from+WG-CIS+to+WG-ISI?preview=/123339734/125468773/Paul%20Knowles%20Re-contributions.pdf
[8] https://kantarainitiative.org/confluence/display/WGISI/Re-contributions+from+WG-CIS+to+WG-ISI?preview=/123339734/125468772/Jan%20Lindquist%20Re-Contributions.pdf

6.4 BIT FIELDS & NOTES

The field(s) below may be represented by single or multiple fields in your application. The overall suggested approach is to be conservative. When reviewing the contents of your dataset against the taxonomy, you should encrypt if the taxonomy might apply, rather than taking a narrow approach. You may find that a field in your dataset might fall within more than one category. That is to be expected as the definitions are somewhat, and intentionally, fuzzy. More precise or prescriptive definitions are the purview of profiles and schemas, where the population of possible field categories can be prescribed or defined more precisely.

#	Field Categories	Notes
1	Names	This includes, but is not restricted to: First Names, Last Names, Full Names, and Entity Names.
2	Physical Address(es)	
3	E-mail Address(es)	
4	Telephone Number(s)	
5	Postal Code(s)	May be included with Physical Address.
6	Personal Software Application Handles	This is a variant on Name. Example sources include Skype, Slack, RocketChat, etc.
7	Profile Pages	
8	Passport Numbers	
9	Social Security Numbers	
10	National Insurance Numbers	
11	Driving License Numbers	
12	Vehicle Registration Numbers	
13	Bank Account Numbers	
14	Financial Institution Card Numbers	This includes but is not restricted to credit or debit card numbers.
15	Personal Identification Numbers (PINs)	
16	Private Keys / Master Keys	
17	Symmetric Keys	
18	Public Keys	
19	Link Secrets	
20	Decentralized Identifiers (DIDs)	See https://w3c.github.io/did-core/

#	Field Categories	Notes
21	Employee Identifiers	This may include identifiers from benefits providers like pension plans.
22	Account Identifiers	
23	Government Identifiers	Numbers, cards or other artefacts issued by a government to a natural person or entity.
24	Membership Identifiers	Examples include but are not restricted to membership in a political party, trade union, fraternal order, survivors groups, or email lists.
25	Institutional Identifiers	Examples include private health care providers, private clubs, and so on.
26	Case Identifiers	Examples include Case ID Numbers, Benefit Plan Participation Identifiers, and so on.
27	User Identifiers	Examples include User IDs, logins, and so on.
28	Passwords	
29	Signatures	Analog or Digital
30	Digital Certificates	Even where a certificate is published and publicly available.
31	Photos	When encrypting files, examine whether the file name should also be encrypted.
32	Videos	When encrypting files, examine whether the file name should also be encrypted.
33	Images	When encrypting files, examine whether the file name should also be encrypted.
34	Vocal Sound Bites	When encrypting files, examine whether the file name should also be encrypted.
35	Dates and timestamps[9]	Examples include Date of Birth[10], transaction dates, and so on.
36	Genetic Identifiers	This includes but is not restricted to chromosomal, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) data.
37	Biometric Identifiers	This includes but is not restricted to voiceprints, iris scans, facial imaging and dactyloscopic (fingerprint) data.
38	Internet Protocol (IP) Addresses	
39	Media Access Control (MAC) Addresses	
40	Service Set Identifiers (SSID)	This includes local WiFi SSIDs.
41	Bluetooth Device Addresses (BD_ADDR)	

[9] Not all captured dates will reveal a person or entity's identity but some will so, if in doubt, encrypt.
[10] In some use cases this can be avoided by using only the Month, or Month/Year of birth, but only if this can be validated.

Blinding Identity Taxonomy

#	Field Categories	Notes
42	Locational Information	This includes Global Positioning System (GPS) or other coordinates, 3-word addresses, and so on.
43	Cookie Browser Identifiers	
44	Radio Frequency Identifiers	
45	IoT Identifiers (incl. smart meter data)	
46	International Mobile Equipment Identity (IMEI)	
47	International Mobile Subscriber Identity (IMSI)	
48	Social media posts and comments	This kind of field may need to be parsed and/or tokenized as part of the blinding process
49	Free-Form Text Fields / Unstructured Data[11]	This kind of field may need to be parsed and/or tokenized as part of the blinding process

Table 3 Blinding Identity Taxonomy

[11] Text which does not have a given structure, nor which is entered in any specific format. Note: All free-form text fields should be encrypted.