ISSN 0719-4110

CIEE
CENTRO DE INVESTIGACIÓN Y ESTUDIOS ESTRATÉGICOS
ANEPE.CL

CUADERNO DE TRABAJO Nº1-2023

GESTIÓN Y ALMACENAMIENTO DE DATOS: DESAFÍOS PARA LA CIBERINTELIGENCIA

Academia Nacional
de Estudios Políticos
y Estratégicos
www.anepe.cl

CUADERNOS DE TRABAJO es una publicación orientada a abordar temas vinculados a la Seguridad y Defensa a fin de contribuir a la formación de opinión en estas materias.

Los cuadernos están principalmente dirigidos a tomadores de decisiones y asesores del ámbito de la Defensa, altos oficiales de las Fuerzas Armadas, académicos y personas relacionadas con la comunidad de defensa en general.

Estos cuadernos son elaborados por investigadores, académicos y colaboradores del CIEE de la ANEPE, pero sus páginas se encuentran abiertas a todos quienes quieran contribuir al pensamiento y debate de estos temas.

Recordamos a los autores que el Cuaderno de Trabajo está comprometido con la publicación de artículos originales e inéditos que difundan conocimiento actualizado en materias de seguridad, defensa y ciencias sociales afines, con el fin de aportar y transferir, con el propósito fundamental de aportar al debate académico múltiples enfoques que enriquezcan el análisis, la reflexión y la interpretación en torno a los temas disciplinares propios de la seguridad, la defensa y las ciencias sociales.

Antes de imprimir este Cuaderno, piense en el medio ambiente.

CUADERNO DE TRABAJO DEL CENTRO DE INVESTIGACIONES Y ESTUDIOS ESTRATÉGICOS es una publicación electrónica del Centro de Investigaciones y Estudios Estratégicos de la Academia Nacional de Estudios Políticos y Estratégicos y está registrada bajo el **ISSN 0719-4110 Cuad. Trab., - Cent. Estud. Estratég.**
Dirección postal: Avda. Eliodoro Yáñez 2760, Providencia, Santiago, Chile.
Sitio Web www.anepe.cl. Teléfonos (+56 2) 2598 1000, correo electrónico ciee@anepe.cl

GESTIÓN Y ALMACENAMIENTO DE DATOS: DESAFÍOS PARA LA CIBERINTELIGENCIA

Cristian Barria Huidobro*

Resumen:

A medida que la información adquiere un papel cada vez más protagónico en la relación sociedad-tecnología, aumenta la necesidad de generar, procesar y almacenar datos en mayor cantidad y con mayor efectividad. Pero con este aumento de datos necesarios para el funcionamiento de la sociedad moderna, también aumenta la cantidad de datos erróneos, falsos, caducados y/o irrelevantes. Esta "basura" de datos genera dificultades en varios niveles: desde ralentizar los tiempos de consulta a las bases de datos, hasta aumentar los costos de almacenamiento digital, por mencionar algunos casos. Este estudio aborda las dificultades que este fenómeno plantea para el campo de la ciberinteligencia, analizando algunas soluciones observadas en diversas organizaciones, y proyectando posibles caminos futuros en torno a este tema.

Palabras clave: Cementerio de datos, basura de datos, ciberinteligencia.

* PhD. Centro de Investigación en Ciberseguridad (CICS) Universidad Mayor Chile. cristian.barria@umayor.cl ORCID: https://orcid.org/0000-0002-5840-7407

DATA MANAGEMENT AND STORAGE: CHALLENGES FOR CYBER INTELLIGENCE

Abstract:

As information takes on an increasingly leading role in the society-technology relationship, the need to generate, process and store data in greater quantities and with greater effectiveness increases. But with this increase in data demand for the functioning of modern society, the amount of erroneous, false, expired and/or irrelevant data also increases. This data "waste" causes difficulties on various levels: from slowing down query times to databases, to increasing digital storage costs, to name a few cases. This study addresses the problems that this phenomenon poses for the field of cyber intelligence, analyzing some solutions observed in various or-ganizations, and projecting possible future paths regarding this issue.

Key words: Data graveyards; data waste; cyber intelligence.

I. Introducción

> "Si no tienes datos, solo eres una persona más con una opinión"[1]. Andreas Schleicher, director de educación de la OCDE

En el mundo interconectado de hoy se producen diariamente grandes cantidades de datos, a una escala difícilmente comparable con cualquier otro período de la historia. Computadores de escritorio, teléfonos móviles, electrodomésticos inteligentes, portales web de compra, sistemas bancarios, agendas médicas virtuales, y un prácticamente interminable etcétera de fuentes producen datos, los cuales son empleados en distintos tipos de análisis, transformándose en información que apoya la toma de decisiones en numerosos ámbitos.

Justamente, a propósito de la gran cantidad de fuentes que producen datos, el volumen que están alcanzando estos últimos se convierten en una espada de doble filo. Por un lado, contar con más datos referentes a un fenómeno específico que se desea analizar, positivo para producir información útil para la toma de decisiones, pero, por otro lado, ante cantidades masivas de datos puede resultar difícil encontrar aquel conjunto de datos que sea específicamente relevante para lo que se desea estudiar. Adicionalmente, ante la posibilidad de que un dato pueda cobrar mayor importancia en el futuro, surge la necesidad de almacenar el dato para uso posterior.

Podemos comparar la situación de los datos con la de un campo en cuyas tierras se cultivan diferentes tipos de plantaciones. Si el terreno se divide en pocas plantaciones, la gestión de lo producido se mantiene ordenada. De igual forma, si el consumo de los alimentos producidos es proporcional a lo que genera este campo ficticio, no hay problemas de sobreabastecimiento ni desabastecimiento. Y

[1] DESMURGET, Michel. La fábrica de cretinos. 2020. Ed. Ariel p. 15

ante la solicitud de un alimento en particular, es posible obtener rápidamente el producto desde la plantación respectiva.

Ahora imaginemos que el mismo equipo agricultor tiene que trabajar cien veces más plantaciones que las del ejemplo anterior, con el triple de variedades de alimentos. La capacidad para poder trabajar todos los cultivos y producir cada alimento se verá fuertemente diezmada. De igual forma, la producción será tal que se requerirá de grandes almacenes para poder asegurar que lo producido no se pierda y pueda alcanzar a futuros consumidores. Y ante una solicitud de un alimento específico, el esfuerzo necesario para identificar una plantación específica y obtener el alimento en cuestión se vuelve mucho mayor.

Podemos pensar en cada tipo de alimento como un dato, mientras que las plantaciones son un conjunto de fuentes de datos específicas. Por ejemplo, los computadores de un hospital, así como su sitio web, su Intranet para médicos y sus bases de datos de clientes representan diferentes fuentes de datos, pero aquellos datos producidos están directa o indirectamente relacionados con el rubro de la salud. Similarmente, en una empresa de logística, los teléfonos celulares de los transportistas tendrán diversos datos sobre sus rutas, las entregas, los paquetes que transportan, geolocalización, entre otros, como también los programas de control de bodegaje que emplea la misma empresa, tendrán otros datos relacionados con el negocio. Cada conjunto de fuentes aportan datos relevantes para un ámbito específico, o a veces para diversos ámbitos.

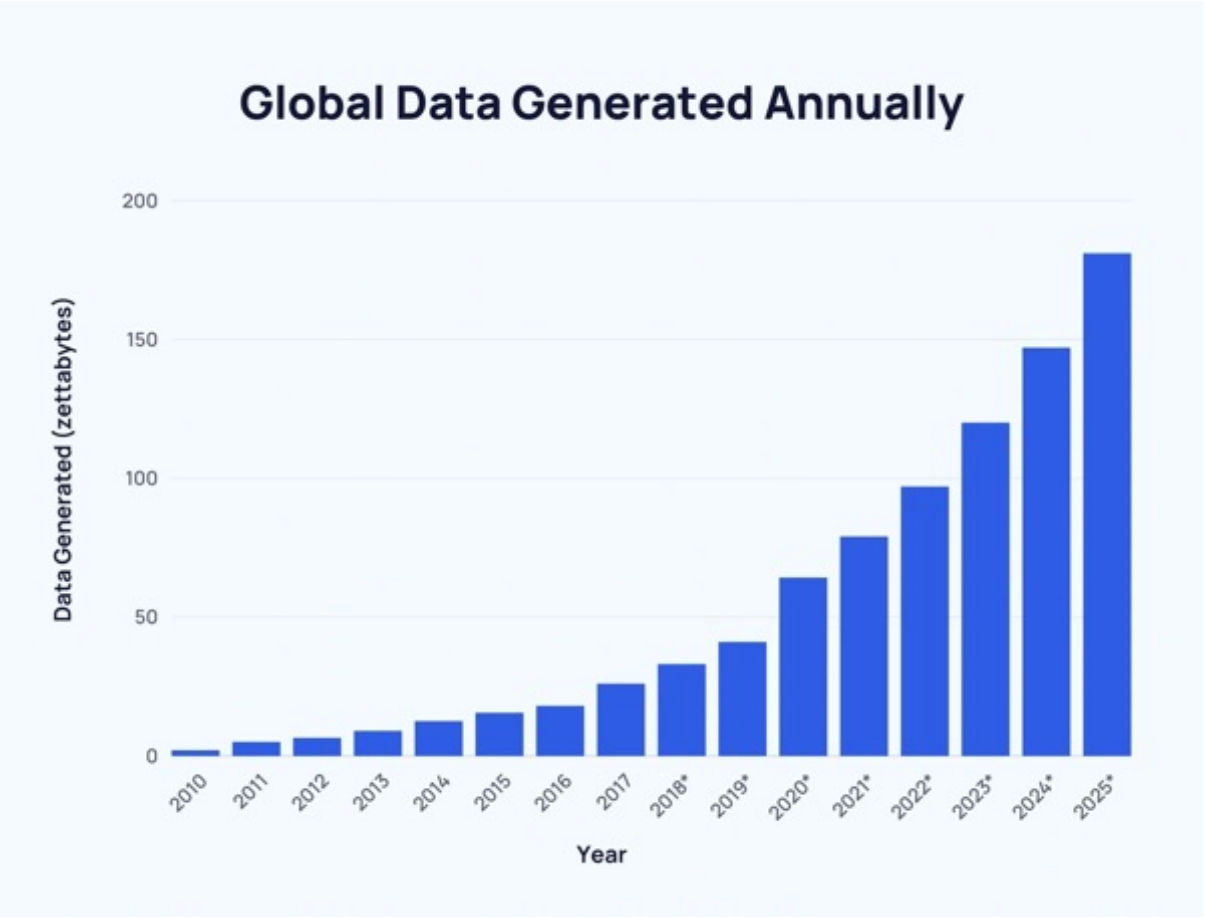
Si llevamos esta idea a la cotidianeidad de cualquier persona, resulta fácil notar que producimos muchos datos diferentes durante el día, e incluso durante la noche, en ciertos casos. Entonces extrapolando este hecho a todo un barrio, a toda una empresa, a toda una ciudad y a todo un país, podemos observar una realidad muy simple: producimos muchos datos todo el tiempo. Sin embargo, puede resultar difícil dimensionar esa cantidad de datos generados.

Al presente año (2023), se estima que cada día se generan 328,77 millones de terabytes diariamente, que equivalen a 328,77 quintillones de bytes. Esto, para ayudarnos a imaginar cuán grande es ese número, equivale a lo siguiente:

328.770.000.000.000.000.000

Y de acuerdo con la tendencia de los últimos años, se estima que esta cantidad seguirá aumentando.

Figura 1: Datos generados anualmente (en zettabytes)[2].



Fuente: NCSI. 2022. National Cybersecurity Index NCSI. ncsi.ega. [En línea] e-Governance Academy Foundation. https://ncsi.ega.ee/ncsi-index/?order=rank&type=c

[2] TAYLOR, P. (2023, august 22). Data Growth Worldwide 2010-2025. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

En la imagen anterior se proyectan valores en zettabytes (ZB), donde cada ZB equivale a 10E21 bytes. Puede ser más simple comparar esta escala de forma gráfica:

Figura 2: Escala visual entre zettabytes, exabytes, petabytes, terabytes y gigabytes[3].

[3] SHINICHI. (2010, september 17). TechNewsDaily. Kushima. http://www.kushima.org/?p=812

En cualquier caso, lo cierto es que estamos produciendo diariamente cantidades gigantescas de datos, lo cual presenta diversas oportunidades, pero también desafíos para el mundo moderno.

Tras la presente Introducción, el presente documento profundiza los conceptos asociados al *Big Data* en la Sección II, continuando con la noción de Ciberinteligencia en la Sección III. Luego en la Sección IV se plantea el marco general de los denominados Datos basura, aspecto profundizado en las secciones V, VI y VII al abordar los conceptos de datos sucios, datos falsos y datos no procesados, respectivamente. Con dichas definiciones individualizadas, se exploran las ideas principales que construyen la noción de cementerio de datos, en la sección VIII. De dicho análisis se desprende la Sección IX, donde se explora el caso particular del acaparamiento de datos. Finalmente, la sección X resume las principales reflexiones obtenidas del estudio, proyectando algunos posibles elementos que puedan conformar una discusión futura sobre la materia.

II. Big Data

En este contexto de grandes cantidades de datos, surge un nuevo concepto en lo que se refiere al procesamiento y análisis de datos: *El Big Data*. Este término se refiere simplemente a conjuntos de datos cuyo volumen es tan grande y tan compleja su variedad, que su tratamiento resulta difícil o inviable con mecanismos tradicionales de procesamiento de datos.

A nivel general, se considera que hablamos de Big Data cuando se encuentran presentes las "Tres V": Volumen, Variedad y Velocidad[4].

- **Volumen:** Puntualmente grandes volúmenes de datos, en diversos ambientes.
- **Variedad:** Una amplia variedad de tipos de datos almacenados.
- **Velocidad**: La alta velocidad a la cual la data es (o debe ser) generada, recolectada y procesada.

Posteriormente se adicionaron otras "V" al paradigma: Valor y Veracidad[5].

- **Valor:** El valor que los datos pueden entregar a la organización, aspecto estrechamente relacionado con qué puede dicha organización hacer con los datos obtenidos.
- **Veracidad:** Esencialmente corresponde al nivel de confianza que existe en la data recolectada, debido a su calidad y exactitud.

Finalmente, se incorporaron dos "V" más: Viabilidad y Visualización[6].

- **Viabilidad:** Capacidad para hacer uso eficaz de los volúmenes de datos.
- **Visualización:** Visualización eficiente de los datos, facilitando la identificación de patrones y tendencias en la información analizada.

Otros autores han esgrimido otras variables como características del *Big Data*, pero en lo concreto es posible establecer una noción más o menos común respecto de lo que representa este concepto. Si volvemos al ejemplo del campo que describimos en la Introducción, el segundo escenario se corresponde con lo que sería *Big Data*: muchos cultivos (grandes volúmenes de datos) que producen numerosos alimentos diferentes (alta variedad de datos) y en grandes cantidades, forzando una labor de cosecha y almacenamiento rápido de los mismos (gran velocidad de generación y recolección de datos).

[4] BOTELHO, B., and BIGELOW, S. J. (2022, january 5). What is Big Data and why is it important?. Data Management. https://www.techtarget.com/searchdatamanagement/definition/big-data

[5] GILLIS, A. S. (2021, march 24). The 5 V's of big data. Data Management. https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data

[6] Instituto de Ingeniería del Conocimiento. (2016, november 29). #Infografía con las 7 vs De #bigdata. https://www.iic.uam.es/innovacion/big-data-infografia-7-v/

El mundo del *Big Data* ha dado pie para el surgimiento de diversas tecnologías y técnicas para el manejo y procesamiento de estos masivos volúmenes de datos, permitiendo así que diversos rubros puedan extraer información relevante para sus necesidades. En nuestro caso puntual, nos interesa el caso de la ciberinteligencia que abordaremos a continuación.

III. Ciberinteligencia

A diferencia de la noción tradicional de Inteligencia, donde existe un enfoque definitorio más o menos establecido, la ciberInteligencia adolece de una problemática ontológica que es común a casi todos los términos asociados al mundo "ciber": la inexistencia de un consenso generalizado y formalmente establecido que provea de una definición concreta. En otras palabras, la definición puntual de ciberinteligencia varía dependiendo de la fuente consultada, a pesar de que puedan existir algunas ideas comunes respecto de qué es lo que significa este término[7].

"Para efectos del presente trabajo, entenderemos ciberinteligencia (CyberINT) como el conjunto estructurado de actividades investigativas orientadas a la recopilación y procesamiento y análisis de información respecto de un objetivo en el ciberespacio."

Para efectos del presente trabajo, entenderemos ciberinteligencia (CyberINT) como el conjunto estructurado de actividades investigativas orientadas a la recopilación y procesamiento y análisis de información respecto de un objetivo en el ciberespacio. Este objetivo puede ser parte del ciberespacio (por ejemplo, una red de equipos) o servirse del ciberespacio mediante acciones digitales (por ejemplo, una persona). Si bien este "objetivo" de estudio de la CyberINT suele ser algún tipo de ciberamenaza, puede también abordar objetivos no necesariamente hostiles.

Por ejemplo, puede ser de utilidad para la organización contar con esta información procesada respecto de una unidad que haya sido recientemente objetivo de intentos de phishing, para determinar el perfil de colaboradores que están bajo la lupa de potenciales atacantes, determinar sus interacciones con entidades externas, entre otros puntos de interés. Esto puede se puede realizar en paralelo a operaciones de ciberinteligencia contra la posible amenaza que sea sospechosa de originar los intentos de phishing, ya que obtenemos así un doble perfilamiento: uno de los atacantes y otro del grupo humano que estos atacantes consideraron como "de interés" para sus intentos.

Dado lo anterior, podemos determinar rápidamente los posibles beneficios de contar con grandes fuentes de datos para analizar. Existiendo los recursos disponibles para realizar operaciones de CyberINT podríamos evaluar tendencias asociadas a las principales ciberamenazas que puedan estar interesadas en atacar a nuestras organizaciones, determinar vectores de ataques comunes, planificar medidas preventivas, desplegar mecanismos de protección, establecer procedimientos de recuperación ante desastres, entre otras opciones.

Una unidad de CyberINT bien nutrida de datos puede apoyar de manera sustancial la toma de decisiones preventivas y correctivas que beneficien a la organización de mejor manera, ya sea mediante la profundización del conocimiento existente ante posibles adversarios, como también mediante un mejor entendimiento de las superficies digitales o humanas propias y sus probables puntos vulnerables.

[7] BONFANTI, Matteo E. Cyber Intelligence: In pursuit of a better understanding for an emerging practice. Cyber, Intelligence, and Security, 2018, vol. 2, no 1, pp. 105-121.

No obstante, el interés puntual por realizar un adecuado perfilamiento de ciberamenazas ha motivado el desarrollo de metodologías principalmente pensadas en abordar estos escenarios, a pesar de que puedan ser empleadas en otro tipo de operaciones de CyberINT que no estén enfocadas en una ciberamenaza propiamente tal. Uno de estos marcos de trabajo populares es el denominado ciclo de vida de la CyberINT de amenazas (*Cyber Threat Intelligence Lifecycle*), el cual consta de seis etapas fundamentales: Dirección, recolección, procesamiento, análisis, diseminación y retroalimentación.

Figura 3: Cyber Threat Intelligence Lifecycle[8].

Fuente: Elaboración propia

[8] Fortra Agari. Cyber threat intelligence: How to stay ahead of threats. Agari. (2021, may 18). https://www.agari.com/blog/what-is-cyber-threat-intelligence

1. **Dirección:** Definición de la información que nos interesa recopilar para analizar el objetivo en cuestión, estableciendo objetivos de manera acorde.
2. **Recolección**: Obtención de la evidencia física o digital de interés.
3. **Procesamiento**: Como su nombre lo indica, esta etapa involucra tomar los datos brutos recopilados y convertirlos en otras formas de datos que sean aptas para su evaluación.
4. **Análisis:** Estudio profundo de los antecedentes recopilados e integración de posibles nuevos datos que hubieren surgido durante las actividades anteriores.
5. **Diseminación:** Entrega de resultados a los tomadores de decisiones correspondientes.
6. **Retroalimentación:** Con posterioridad a las acciones que se hayan tomado gracias a la información obtenida, se evalúan los resultados y se emplean en una nueva iteración del ciclo, aportando antecedentes a la etapa de dirección.

Si nos situamos en nuestro escenario ficticio anterior, donde se disponen de grandes cantidades de datos para análisis y los recursos necesarios para llevar a cabo las operaciones de CyberINT, podemos formular algunas preguntas que nos darán luces de las posibles problemáticas que podría presentar esta situación, a pesar de que inicialmente sea algo positivo.

- ¿Qué pasa si en estos grandes volúmenes de datos encontramos casos que no sirven para nuestro análisis?
- ¿Podemos validar si hay datos falsos, erróneos o desactualizados?
- ¿Cuánto almacenamiento de datos necesitaremos en el corto, mediano y largo plazo para poder realizar nuevos análisis?

IV. Los datos basura

> "Qué pasa con la POSVERDAD, con esa información desbordada de mentiras. En otros tiempos se les llamaba Herejías"[9].

Los datos basura, también conocidos como "j*unk data*", son información inútil o irrelevante que se acumula en un conjunto de datos. Estos datos no aportan ningún valor y, en cambio, ocupan espacio y pueden dificultar el análisis efectivo. El origen de los datos basura puede deberse a errores en la recolección, almacenamiento o transferencia de información. En el contexto de Big Data es crucial limpiar y filtrar estos datos para garantizar que no contaminen los resultados del análisis.

"Los datos basura, también conocidos como "*junk data*", son información inútil o irrelevante que se acumula en un conjunto de datos. Estos datos no aportan ningún valor y, en cambio, ocupan espacio y pueden dificultar el análisis efectivo."

Estos datos basura pueden abarcar diversos casos; desde datos generados de manera innecesaria por sistemas redundantes (o incluso por malos procedimientos), hasta datos falsos producidos por actores hostiles con la finalidad de obstruir nuestras operaciones de CyberINT. En otros casos, puede ocurrir que por una labor deficiente durante la etapa de Dirección se termine recopilando datos que a pesar de haber sido inicialmente catalogados como relevantes para la operación, al final resultan ser completamente inútiles. En ese sentido, es responsabilidad del equipo involucrado en el CyberINT poder identificar oportunamente este tipo de errores y rectificar según corresponda.

[9] BARRICO, Alssandro. The Game. Ed. Anagrama. 2019. p. 284

También existe la posibilidad de una captación excesiva de datos. Esto se refiere a una obtención de datos en cantidades mayores a las necesarias para realizar análisis. Un ejemplo podría ser analizar miles de correos electrónicos para determinar un ataque de phishing, cuando usualmente contando con un único correo de muestra se puede determinar un eventual origen fraudulento.

Además, podemos encontrar situaciones en las que ante un conjunto determinado de datos producidos por fuentes determinadas, en una operación de CyberINT se determine que los conjuntos de datos A, B y C son útiles, mientras que en otra operación diferente, solo los conjuntos A y C ameriten estudio. En este escenario, el conjunto B se vuelve técnicamente un grupo de datos basura por la naturaleza de la operación en curso, no por un error o una acción mal intencionada.

Una apropiada limpieza de datos es de suma importancia para la CyberINT, ya que, a diferencia de nuestro caso ficticio ideal, en el día a día no dispondremos de recursos ilimitados para ignorar la carga que significa contar con estos elementos que finalmente no aportan a la investigación. Dependiendo de la urgencia de la situación, el tiempo extra que adicione la gestión de estos datos basura puede significar que los tomadores de decisiones no cuenten con la información necesaria a tiempo, pudiendo retrasar la ejecución de acciones críticas.

En su libro, Bartlet describe a un grupo llamado GNA (Asociación Nacional de Maulladores GNA en inglés) que tenía como objetivo sembrar incertidumbre en Internet, empleando la inundación de datos basura, "*a menudo inundaban sitios con basura; llenaban las funciones de chat con sinsentidos, tal como lo hacían los maulladores una década antes, y hackeaban otros sitios populares*"[10].

Este tipo de ataque se caracteriza por el envío masivo y repetido de información inútil, sin sentido o maliciosa a través de diferentes medios en línea, como sitios web, foros, funciones de chat y redes sociales. El objetivo principal de este tipo de ataque es saturar los sistemas y recursos en línea, generando un caos digital y dificultando el funcionamiento normal de las plataformas afectadas. Este tipo de comportamiento malicioso y disruptivo puede causar serios problemas para los sitios web y sus usuarios, ya que puede sobrecargar los servidores, dificultar la interacción entre usuarios legítimos y causar pérdidas económicas y de reputación.

"Dependiendo de la urgencia de la situación, el tiempo extra que adicione la gestión de estos datos basura puede significar que los tomadores de decisiones no cuenten con la información necesaria a tiempo, pudiendo retrasar la ejecución de acciones críticas."

El propósito detrás de este tipo de ataques puede variar desde fines de entretenimiento y la búsqueda de notoriedad hasta acciones más nefastas, como la desinformación, la difamación o el sabotaje de sitios y servicios en línea.

Es importante destacar que la lucha contra la inundación de datos basura y otras formas de ataques cibernéticos es un desafío constante para la seguridad en línea. Los administradores de sitios web y plataformas, así como los usuarios, deben estar alertas y tomar medidas para protegerse contra estos ataques, tales como implementar medidas de seguridad, utilizar software actualizado y aplicar buenas prácticas de seguridad en línea.

[10] BARTLETT, J. (2017). In La Red Oculta (p. 18). essay, Planeta Publishing.´

Ahora bien, el ejemplo recién mencionado corresponde a datos basura producidos deliberadamente por un adversario pero, como explicamos anteriormente, no siempre existe este tipo de intencionalidad. En esta nueva vereda encontramos el caso de los denominados "datos sucios", los cuales dicen relación con datos de baja calidad.

V. Los datos sucios

En su nivel más básico, un dato sucio es un dato incorrecto, como podría ser un nombre mal escrito, un registro mal copiado, fechas sin sentido, entre otros. Respecto de este tema, Susan Walsh menciona lo siguiente[11]:

> "Los datos sucios son un problema. En cada organización, sin importar su tamaño o ubicación, se hablará de problemas de calidad de datos. Sin embargo, rara vez se mencionarán las consecuencias de esto, ya que las personas o empresas no quieren admitir sus fallos. Podríamos estar hablando de millones de libras o dólares perdidos en nuevas tecnologías, semanas o meses gastados corrigiendo errores debido a datos incorrectos, posibles pérdidas de empleo o incluso algo peor.
>
> Además de eso, constantemente escuchamos que los científicos de datos dedican entre un 40% y un 80% de su tiempo a limpiar o manipular datos. ¿Por qué ocurre esto? Bueno, creo que se debe a su ineficiencia e inexperiencia en este aspecto. Puedes pensar: "¿Pero son científicos de datos!" Lamentablemente, eso no lo soluciona todo. La limpieza de datos rara vez se aborda en estudios académicos u otros cursos; el enfoque siempre está en los aspectos técnicos del rol, aunque irónicamente, no pueden realizar ninguna de esas tareas sin tener primero datos limpios.
>
> Aunque la limpieza de datos es una de las partes más vitales de todo el proceso al trabajar con datos, a menudo se pasa por alto porque se asume que las personas ya saben cómo hacerlo o se considera demasiado trivial o no lo suficientemente importante como para invertir tiempo o recursos. Esto no se limita solo a la ciencia de datos".

De esta forma, Welsh evidencia que los desafíos que la basura de datos presenta para la CyberINT también abarca factores internos en la organización, la cual agrega una dimensión diferente que los tomadores de decisiones deben considerar para asegurar operaciones fluidas y efectivas: la auditoría interna y la mejora continua en procesos y procedimientos.

VI. Los datos falsos

> "La buena espada de la verdad sólo se mantiene afilada si se la pone a prueba constantemente contra las hachas y las porras de la falsedad."[12]

Si los maulladores de Bartlett representan un claro ejemplo de basura de datos producida de forma externa a la institución, los científicos de datos de Welsh son la otra cara de la moneda, donde la basura se produce por desprolijidades internas. En este contraste de blanco y negro, existen diversas escalas de grises.

Tal es el caso de la "*Fake Data*". Como su nombre lo indica, se trata de aquellos datos falsos, resultado de errores humanos, problemas técnicos o deliberadamente manipulados o creados con el propósito de engañar o distorsionar el análisis. Estos datos pueden surgir de diversas fuentes internas o externas, y su presencia puede llevar a conclusiones incorrectas y decisiones erróneas.

[11] WALSH, Susan. Between the spreadsheets: classifying and fixing dirty data. Facet Publishing, 2021.
[12] Op. Cit. DESMUURGET. 2021. p. 115.

Más aún, si estamos ante la misión de analizar amplios y complejos conjuntos de datos (como en el caso del *Big Data*), la intrusión de datos falsos desde diversas fuentes puede ir en detrimento de las operaciones de CyberINT.

VII. Los Datos No Procesados

En una escala más inofensiva, pero que de todos modos representa un desafío importante para la ciberinteligencia, encontramos el problema de los datos no procesados, también conocidos como datos en bruto o datos sin procesar. Estos corresponden a aquellos en su forma original, tal como se recopila de fuentes diversas, sin haber sido sometidos a ningún tipo de manipulación, análisis o transformación. Estos datos pueden presentarse en diferentes formatos, como texto sin estructurar, imágenes, audio, video o datos numéricos sin formato.

La naturaleza de los datos no procesados puede variar según la fuente de donde provengan. Por ejemplo:

- **Datos textuales:** Incluyen registros de texto sin formato como comentarios de clientes, transcripciones de entrevistas, correos electrónicos, informes o documentos no estructurados.
- **Datos numéricos**: Representan información en forma de números sin procesar, como registros de ventas, mediciones de sensores, datos financieros sin transformar, entre otros.
- **Datos multimedia:** Comprenden imágenes, audio y video en su formato original sin ningún tipo de edición o codificación específica.

Estos datos en bruto pueden ser difíciles de interpretar y analizar directamente debido a su falta de estructura o formato coherente. Por lo tanto, es común que los datos no procesados se sometan a un proceso de preparación y limpieza antes de ser utilizados para fines analíticos o de toma de decisiones. Este proceso incluye actividades como estructuración de datos textuales, la eliminación de datos duplicados o incompletos, la normalización de formatos y la conversión de datos a una estructura coherente y comprensible, entre otros.

Una vez que los datos no procesados se transforman en datos procesados, es posible realizar análisis, visualizaciones, estadísticas y aplicar algoritmos de aprendizaje automático capaces de extraer información valiosa y conocimiento útil para diversas aplicaciones en campos como la inteligencia empresarial, la investigación científica, la medicina, la ingeniería, entre otros.

Figura 4: Tipos de datos no procesados.

Fuente: Elaboración propia.

Si bien este tipo de datos no necesariamente constituirán basura cuando sean procesados, lo cierto es que estas labores de preparación para su análisis son fundamentales para las operaciones de CyberINT, ya que sin ellas no es posible estudiar los datos y producir el conocimiento deseado.

Sin ir más lejos, el propio ciclo de vida de CyberINT de amenazas que discutimos previamente, establece una etapa específicamente enfocada en el tratamiento de los datos brutos para su posterior análisis. Podemos decir incluso que en el ciclo se considera de forma explícita la existencia de —al menos—este tipo de basura de datos.

Todos estos puntos que hemos revisado abordan algunas de las preguntas que nos habíamos planteado, pero queda una sin abordar: aquella respecto del almacenamiento de datos. De aquí surgen ciertos conceptos importantes que se discuten a continuación.

VIII. Los cementerios de datos

> "Cuando las cosas se vuelven demasiado complicadas, a veces tiene sentido parar y preguntarse: ¿He planteado la pregunta correcta?"[13]

La idea del desecho y de los "Cementerios de datos" se centra en comparar la acumulación masiva de información no deseada o innecesaria en el mundo digital con el problema de los desechos en la vida cotidiana. A través de estas analogías, se resaltan los peligros y las amenazas que esta acumulación puede representar para nuestra seguridad digital y privacidad, así como para la eficiencia de nuestras plataformas y sistemas en línea. También enfatiza la necesidad

[13] BOMBIERI, E. "Prime territory". En: The Sciences. DU SAUTOY, M. La música de los números primos. p. 37

de adoptar un enfoque activo para gestionar y limpiar adecuadamente nuestros datos, al igual que las civilizaciones antiguas encontraron soluciones creativas para abordar el problema de los desechos en su desarrollo urbano.

La analogía destaca cómo el mundo digital se ha convertido en un vasto cementerio de datos que alberga información obsoleta, innecesaria o dañina, de manera similar a cómo las heces pueden contener bacterias y virus peligrosos. Al igual que una acumulación descontrolada de desechos puede amenazar la salud, la acumulación descontrolada de datos en línea puede amenazar nuestra seguridad digital y privacidad. Así como la científica Valerie Curtis advirtió sobre los peligros de las heces y llamó a la acción para mejorar nuestros sistemas de saneamiento, también se deben tomar medidas proactivas para gestionar y limpiar adecuadamente nuestros datos para proteger nuestra experiencia digital.

“Podemos afirmar que estamos frente a un cementerio de datos cuando se ha producido una acumulación de datos cuyo almacenamiento no obedece a una necesidad de uso razonable, sino que a fallos o descuidos en lo que respecta a su eliminación.”

Podemos afirmar que estamos frente a un cementerio de datos cuando se ha producido una acumulación de datos cuyo almacenamiento no obedece a una necesidad de uso razonable, sino que a fallos o descuidos en lo que respecta a su eliminación.

Esto puede responder a diferentes causales. Por ejemplo, si la organización no establece políticas de ciclo de vida de los datos, éstos se irán acumulando sin restricciones a lo largo del tiempo, a pesar de que su utilidad haya expirado. Recordemos que si bien es cierto que disponer de datos antiguos nos aporta valor al permitir analizar comportamientos a lo largo del tiempo y determinar tendencias —que a su vez pueden emplearse para estimar escenarios futuros—, cuando un conjunto de datos se vuelve obsoleto, ya no puede apoyar a la toma de decisiones de la organización, por lo que su almacenamiento solo supone costos adicionales en la forma de discos duros, cintas de respaldo, archivos en la nube, entre otros.

La arista normativa

Adicionalmente, dependiendo del tipo de datos almacenados, un cementerio de datos puede implicar problemas normativos. Tal es el caso cuando requerimos almacenar datos de personas, ya sean colaboradores, clientes, usuarios u otros, ya que este tipo de datos suelen estar sujetos a leyes locales enfocadas en la protección de la privacidad personal. A nivel personal, en este aspecto destaca el Reglamento General de Protección de Datos (RGPD o GDPR por sus siglas en inglés), una regulación europea justamente centrada en aspectos de privacidad de la información, cuyo alcance abarca a todos los países de la Unión Europea (UE) y a aquellas instituciones que, aunque operen fuera del territorio de la UE, trabajen con datos de ciudadanos europeos.

En el contexto actual de alta conectividad e interacciones tanto personales como de negocios fuera de las fronteras nacionales, es habitual que distintas instituciones se encuentren dentro del alcance del RGPD, lo cual las obliga a implementar controles de cumplimiento normativo para no transgredir este reglamento.

Por lo anterior, la existencia de cementerios de datos supone un riesgo de incumplimiento, ya que esto transgrede el principio de que los

datos personales solo pueden ser almacenados para el fin específico que motivó inicialmente su recolección, y únicamente durante el periodo de tiempo en el cual este almacenamiento sea necesario para cumplir dicho fin.

Podemos ir más allá y proyectar posibles consecuencias de esto. Si la organización posee un cementerio de datos y sufre una fuga producto de un ciberataque, dependiendo de cuán público sea el hecho —recordemos que es práctica relativamente común entre los ciberdelincuentes el difundir las fugas para atraer a potenciales compradores de datos— el impacto reputacional para la institución puede ser muy importante, además de los costos económicos adicionales que va a requerir el remediar los problemas que hicieron posible esa fuga, sumado a la revisión de activos, procedimientos y otros puntos.

"Un cementerio de datos se trata, entonces, de una fuente de diversos problemas, desde el ya mencionado incremento de costo de almacenamiento, hasta los ejemplos recién descritos referentes al ámbito normativo."

Si alguno de los repositorios de datos afectados estaba sujeto al RGPD, entonces aparte de todo lo ya ocurrido, la institución tendrá que someterse a una auditoría, la cual arrojará la existencia de un incumplimiento a la norma, pudiendo transformarse en sanciones de distinto calibre.

Un cementerio de datos se trata, entonces, de una fuente de diversos problemas, desde el ya mencionado incremento de costo de almacenamiento, hasta los ejemplos recién descritos referentes al ámbito normativo. Cabe destacar, además, que a medida que crecen estos repositorios se va requiriendo de más recursos para poder monitorear y proteger a estos activos, haciendo que la superficie a proteger se vuelva cada vez mayor y más costosa. Respecto de esta relación entre seguridad y costos, la firma *Data Privacy Manager*[14] menciona lo siguiente:

"La seguridad de los datos es uno de los aspectos más costosos que debe considerar al almacenar grandes cantidades de datos, y es algo que no se puede ignorar. La seguridad de los datos puede tener muchas capas, comenzando con el almacenamiento e incluyendo el cifrado. Se implementa para bloquear el acceso de terceros a los datos.

Para garantizar la seguridad completa de los datos, se necesitará un plan de seguridad de datos estricto y un equipo que trabaje las 24 horas del día, los 7 días de la semana. Debe hacer que el equipo se adhiera a las mejores prácticas de seguridad. Si se está externalizando el almacenamiento de datos, debe elegir los mejores socios en los que pueda confiar.

Además, debe estar atento y proteger sus datos de las amenazas emergentes, con especial atención a los ciberataques. Su instalación de almacenamiento de datos necesita usar un sistema robusto y tener medidas de seguridad física 24/7/365. De lo contrario, corre el riesgo de sufrir graves violaciones de datos.

No es fácil conseguir la seguridad absoluta, pero hay que poner todas las medidas. Es posible que el software de seguridad, como el antimalware y otros sistemas de seguridad, no sea una garantía para la seguridad de los datos porque las filtraciones de datos también pueden provenir de un trabajo interno. A veces, un atacante puede hacerse pasar por un cliente y ejecutar una carga útil de día cero en su servidor".

Vida útil y depreciación de los datos

Esta noción que esgrime el RGPD respecto de que los datos solo pueden ser almacenados durante un período de tiempo específico, apunta a dos conceptos fundamentales que, si bien comparten un símil con el mundo de los activos físicos, en el contexto digital poseen características diferentes: la vida útil y la depreciación de los datos.

[14] DPM. (2022, march 15). Data graveyards: Challenges and risks. Data Privacy Manager. https://dataprivacymanager.net/data-graveyards-challenges-and-risks/

A diferencia de lo que ocurre con los activos físicos como las maquinarias y el equipamiento, las nociones de vida útil y depreciación se emplean para describir el valor y la utilidad de los datos a lo largo del tiempo.

Como su nombre lo indica, la vida útil de los datos, hace referencia al periodo de tiempo en el cual un dato (o un conjunto de ellos) es considerado como "útil" para la organización que los desea conservar. Al igual que en el caso de otros activos intangibles como la propiedad intelectual o las patentes, esta utilidad percibida respecto de los datos puede cambiar con el pasar del tiempo, pudiendo estos cambios ser influenciados por factores como la evolución de la tecnología, cambios en las regulaciones, decisiones comerciales y la naturaleza cambiante de las necesidades de los usuarios.

Por otro lado, cuando hablamos de una depreciación de los datos estamos ante una noción que aborda la disminución del valor percibido de un dato a lo largo del tiempo. Podemos en este caso observar elementos habituales que influyen en la depreciación de un dato: Obsolescencia, pérdida de relevancia, desgaste de la calidad, y cambios en el contexto.

1.**Obsolescencia:** Cuando el ingreso de nuevos datos actualiza nuestro conocimiento de un fenómeno de interés, suele ocurrir que aquellos datos que hacen referencia a mediciones anteriores del mismo fenómeno pasen a perder vigencia, o dicho de otro modo, se vuelven obsoletos. Un dato obsoleto puede eventualmente ser utilizado para efectos de comparaciones históricas, pero de todos modos es un dato cuyo valor es mucho menor que cuando fue recopilado inicialmente.
2. **Pérdida de relevancia**: Cuando el objetivo que la organización desea lograr cambia, los datos puntuales que sean percibidos por ella como "útiles" también pueden cambiar. Por ejemplo, si una entidad minera centró el foco de sus operaciones en la extracción de plata, pero por variaciones del mercado decide volcar todos los esfuerzos a la explotación de oro, entonces los datos de suelo relacionados con la plata pierden relevancia, reduciendo su valor.
3. **Desgaste de la calidad:** Con el tiempo, los datos pueden perder calidad debido a errores, corrupción o falta de actualización. Datos poco confiables o desactualizados pueden tener un impacto negativo en las decisiones comerciales, y se consideran menos valiosos que datos de mejor calidad.
4. **Cambios en el contexto:** Los datos están vinculados al contexto en el que se recopilaron y analizaron. A medida que el contexto cambia, es posible que los datos ya no sean aplicables o precisos. Estos cambios de contexto pueden ser a nivel mercantil, político, social, tecnológico o incluso cultural.

Figura 5: Dimensiones de la depreciación de los datos.

Fuente: Elaboración propia.

Es importante reconocer que no todos los datos siguen el mismo patrón de depreciación. Algunos conjuntos de datos pueden mantener su valor durante períodos más largos debido a su naturaleza fundamental o su relevancia continua.

Estos dos macroconceptos exigen a las organizaciones una apropiada gestión de sus datos, estableciendo criterios para definir los ciclos de vida de los datos, controles de evaluación y actualización de datos tanto a nivel de recolección como de almacenamiento y eliminación, además de ir evaluando con una regularidad razonable —lo que dependerá de la naturaleza de cada data y las necesidades organizacionales— cuán alineados están estos conjuntos de datos con los objetivos de la institución, velando así por su relevancia y calidad.

Una gestión integral de estos repositorios y almacenes digitales ayuda a evitar la formación de cementerios de datos, al combatir —entre otras cosas— lo que se conoce como el acaparamiento de datos o "*Data Hoarding*".

IX. Acumulación de Datos (Data Hoarding)

El acaparamiento de datos, o acaparamiento digital, es un fenómeno que se caracteriza por una adquisición de contenidos digitales pero acompañados por una incapacidad para descartar tales contenidos, lo cual puede conducir a una acumulación descontrolada de datos basura, como relatan Sedera y Lokuge[15]. Esta conducta ha despertado creciente interés en el mundo de la investigación, donde diversos autores han analizado los efectos que este acaparamiento digital está ocasionando en diversos ámbitos.

En el año 1997, Marshall abordó el debate ético producido entre científicos, empresarios biotecnológicos y firmas farmacéuticas respecto de quién debe almacenar y controlar los datos de secuencias de ADN. En la investigación se plantea si este tipo de acaparamiento ha significado un retraso en el avance científico en este rubro[16].

[15] SEDERA, D., & LOKUGE, S. (2018, january). Is digital hoarding a mental disorder? Development of a construct for digital hoarding for future IS research. In Proceedings of the 39th International Conference on Information Systems (ICIS 2018). University of Southern Queensland.

[16] MARSHALL, E. (1997). Ethics in science: Is data-hoarding slowing the assault on pathogens?. Science, 275(5301), 777-780.

McKellar, Sillence, Neave y Briggs estudiaron la relación entre la cultura organizacional y las tendencias de acaparamiento de datos observado en los trabajadores, recalcando la importancia de desarrollar políticas organizacionales apropiadas para prevenir y regular el acaparamiento digital[17].

Por su parte, Cynthia y Samantha Gormley analizaron el impacto que este tipo de acaparamiento produce en diferentes dimensiones institucionales: los costos, la productividad y la cultura organizacional[18].

Lo cierto es que a diferencia del acaparamiento de objetos que caracteriza ciertos desórdenes mentales, como el síndrome de Diógenes, el acaparamiento de datos no necesariamente responde a un factor psicopatológico, sino más bien a una compleja combinación de factores humanos y organizacionales. Y respecto de esto último, otra gran diferencia es que el acaparamiento digital justamente puede ocurrir a nivel institucional, a diferencia de Diógenes, que es un caso individual —aunque es cierto que hay casos donde este desorden se presenta en grupos familiares, y no solo en una persona específica—, pasando a ser un reflejo de procesos, procedimientos y cultura organizacional con fallos que incluso pueden pasar desapercibidos.

En cierto sentido, el acaparamiento digital es un catalizador para la creación de cementerios de datos por cuanto este fenómeno transgrede los criterios de eliminación de datos (algo clave para respetar sus ciclos de vida), excediendo así los parámetros que rigen el “almacenamiento razonable” de los mismos.

Por lo anterior, en complemento a las medidas técnicas, administrativas y normativas que se puedan implementar para prevenir la aparición de cementerios de datos, ahora se debe integrar el factor humano y cultural, para poder contar con el impulso de nuestros equipos de personas, quienes son finalmente los que con sus acciones y costumbres implementan lo que de otra forma sólo se quedaría en la teoría escrita dentro de una política o un instructivo.

En síntesis, es factible afirmar que la CyberINT afronta significativos desafíos que pueden ser categorizados en dos aspectos principales: el almacenamiento y la gestión. Estos aspectos encapsulan las particularidades de los datos mencionados previamente y quedan reflejados en la figura 6.

[17] MCKELLAR, K., SILLENCE, E., NEAVE, N., & BRIGGS, P. (2023). Digital accumulation behav-iours and information management in the workplace: exploring the tensions between digital data hoarding, organisational culture and policy. Behaviour & Information Technology, pp1-13.

[18] GORMLEY, C. J., & GORMLEY, S. J. (2012). Data hoarding and information clutter: The impact on cost, life span of data, effectiveness, sharing, productivity, and knowledge management cul-ture. Issues in Information Systems, 13(2), 90-95.

Figura 6: Desafíos para la CyberINT en el almacenamiento y gestión de los datos.

Fuente: Elaboración propia.

X. Consideraciones finales

El mundo de la CyberINT requiere de una gestión integral y constante para poder cumplir su objetivo de apoyar la toma de decisiones mediante la producción de información oportuna y fidedigna. En el corazón de este tipo de operaciones yacen los datos que, tanto por su propia naturaleza como por el efecto de factores internos y externos a la organización, exigen de un cuidadoso escrutinio en todas las etapas del ciclo de vida de una operación de CyberINT, además de apoyarse en buenas prácticas de gestión en otras aristas organizacionales (como la auditoría de procesos y procedimientos).

Al igual que en otras dimensiones relacionadas directa o indirectamente con la Seguridad de la Información, la correcta aplicación de la CyberINT depende de un esfuerzo institucional que va más allá de la operación misma ya que, como vimos a lo largo del presente documento, existen diversos elementos que pueden impactar negativamente al éxito de este tipo de labores de inteligencia.

El mundo interconectado de hoy pareciera seguir la tendencia de una producción cada vez mayor de datos, fomentando los beneficios del Big Data, pero también desafiando a la CyberINT con diversos obstáculos que se manifiestan en las distintas formas que toma el almacenamiento y la gestión de los datos.

Podemos hablar tranquilamente de una problemática que, si no es abordada por las instituciones, eventualmente generará perjuicios tangibles a la organización, incluso pudiendo llegar a daños económicos, pérdidas de datos críticos, o incluso a transgredir las normativas vigentes de protección de datos.

Tal como ocurre con la Ciberseguridad en general, las personas encargadas de dirigir los esfuerzos en materia de CyberINT deben tener la visión necesaria para detectar las principales vulnerabilidades y amenazas que afecten a su institución en esta materia, alineando sus objetivos con los de la organización y transmitiendo a los tomadores de decisiones la

importancia de contar con apoyos transversales que permitan velar por la producción de datos de calidad y mitigar los efectos del almacenamiento y de la gestión de los datos, manteniendo a su vez un ojo vigilante para equilibrar las necesidades razonables de estos aspectos de los datos sin perder control de los ciclos de vida respectivos que permitan eliminar datos "muertos" de manera oportuna y eficiente.

En un contexto de recursos limitados para el manejo de datos que ya exceden la escala de los zettabytes, no cabe duda de que con el pasar de los años cada unidad de CyberINT tendrá que lidiar con volúmenes cada vez mayores de datos altamente complejos, desafío que solo podrá ser abordado con éxito mediante una gestión de datos de forma ordenada, eficiente y cuidadosamente planificada.

Finalmente, es de suma importancia conservar una perspectiva integrada del rol de la CyberINT, ya que no sólo debe interactuar con otras áreas institucionales, sino que también se sitúa dentro de procesos más complejos referentes a la ciberseguridad y la seguridad de la información general. En un trabajo anterior[19] se abordó la naturaleza múltiple del ciberespacio, en la que convergen ámbitos como las ciberoperaciones y la misma CyberINT. Por lo anterior, la CyberINT debe ser capaz de conversar con otros ámbitos institucionales tanto de forma horizontal como vertical, ya que de lo contrario no se podrá agregar valor a los tomadores de decisiones, especialmente en situaciones de urgencia donde la capacidad para determinar rápidamente el mejor curso de acción determina si se puede superar exitosamente una situación de crisis o no.

Si bien el análisis completo del rol de la CyberINT dentro de la seguridad de la información excede al alcance del presente estudio, es importante recordar que conforme a lo recientemente expuesto, CyberINT es un esfuerzo permanente, cíclico y transversal que se ve enfrentado a desafíos altamente complejos, pero que con una planificación responsable y la colaboración institucional, es posible abordar de manera correcta y situar a la organización en un punto donde se pueda responder de manera oportuna a las vicisitudes que el mundo interconectado moderno plantea día a día.

[19] BARRÍA, C. (2019). La dimensión del ciberespacio: una propuesta de ciberseguridad. Cuaderno de trabajo Nº1-2019, Academia Nacional de Estudios Políticos y Estratégicos. (1), 1-21.

BIBLIOGRAFÍA

BARRÍA, C. (2019). La dimensión del ciberespacio: una propuesta de ciberseguridad. Cuaderno de trabajo Nº1-2019, Academia Nacional de Estudios Políticos y Estratégicos. (1), 1-21.

BARTLETT, J. (2017). La Red Oculta. Planeta Publishing.

BONFANTI, M. E. (2018). Cyber Intelligence: In pursuit of a better understanding for an emerging practice. Cyber, Intelligence, and Security, 2(1), 105-121.

BOTELHO, B., and BIGELOW, S. J. (2022, january 5). What is Big Data and why is it important?. Data Management. https://www.techtarget.com/searchdatamanagement/definition/big-data

Cyber threat intelligence: How to stay ahead of threats. Agari. (2021, may 18). https://www.agari.com/blog/what-is-cyber-threat-intelligence

Data graveyards: Challenges and risks – Data Privacy manager. Data Privacy Manager. (2020, August 11). https://dataprivacymanager.net/data-graveyards-challenges-and-risks/

GILLIS, A. S. (2021, march 24). The 5 V's of big data. Data Management. https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data

GORMLEY, C. J., & GORMLEY, S. J. (2012). Data hoarding and information clutter: The impact on cost, life span of data, effectiveness, sharing, productivity, and knowledge management culture. Issues in Information Systems, 13(2), 90-95.

Instituto de Ingeniería del Conocimiento. (2016, november 29). Infografía Big Data: las 7 V. Instituto de Ingeniería del Conocimiento. https://www.iic.uam.es/innovacion/big-data-infografia-7-v/

MARSHALL, E. (1997). Ethics in science: Is data-hoarding slowing the assault on pathogens?. Science, 275(5301), 777-780.

MCKELLAr, K., SILLENCE, E., NEAVE, N., & BRIGGS, P. (2023). Digital accumula-tion behaviours and information management in the workplace: exploring the ten-sions between digital data hoarding, organisational culture and policy. Behaviour & Information Technology, 1-13.

SEDERA, D., & LOKUGE, S. (2018, january). Is digital hoarding a mental disorder? Development of a construct for digital hoarding for future IS research. In Proceed-ings of the 39th International Conference on Information Systems (ICIS 2018). Uni-versity of Southern Queensland.

TAYLOR, P. (2022, september 8). Total Data Volume Worldwide 2010-2025. Statis-ta. https://www.statista.com/statistics/871513/worldwide-data-created/

TechNews Daily. Kushima.org. (2013, march 28). http://www.kushima.org/?p=812

WALSH, S. (2021). Between the spreadsheets: Classifying and fixing dirty data. Facet Publishing.


CENTRO DE INVESTIGACIÓN Y ESTUDIOS ESTRATÉGICOS
ANEPE.CL