ESPACIOS DE DATOS FEDERADOS

DAVID PÉREZ

AEAT

JUAN JIMÉNEZ MORILLAS

Acción contra el Hambre

En 1969 Peter Drucker escribió *The age of discontinuity*, donde consignó el concepto de «sociedad del conocimiento». En el libro, Drucker coloca al conocimiento en el centro de los procesos de generación de riqueza y apuntaba a que lo importante no era tanto la *cantidad* como la *productividad* del conocimiento. Casi medio siglo después, día a día se actualiza esta idea de la mano fundamentalmente de dos grandes familias de tecnologías: el procesamiento de grandes volúmenes de información, en ocasiones llamado BigData, y la Inteligencia Artificial, que permiten la explotación de los datos para convertirlos en información y luego en conocimiento (1) de un modo que era imposible entonces.

En ambos casos el objeto de su actividad, la «materia» sobre la que operan, son los datos. La esencia del BigData es que engloba la familia de técnicas para la gestión y procesamiento de gran cantidad de datos y a gran velocidad. En el caso del aprendizaje automático, se requiere de gran cantidad de datos curados para el entrenamiento de los algoritmos, algoritmos que luego trabajarán a su vez sobre algún tipo de dato.

Así, los datos se han convertido hace ya tiempo en una nueva materia prima, un recurso clave para la economía entendida en un sentido amplio no sólo como generación de riqueza, sino también como la ciencia de la gestión de los recursos. Asumiendo que vivimos en una sociedad del conocimiento, es cada vez más evidente que el sistema económico operante es una verdadera economía del conocimiento (desmaterialización de productos, mercados de intangibles, explotación de datos, toma de decisiones basada en ellos). Del mismo modo que la máquina de vapor cambió los modos de producción y abrió la puerta a la sociedad industrial, el uso y la explotación de los datos ha transformado la sociedad en una verdadera sociedad digital. Hablar por tanto de la importancia del dato a estas alturas es casi un lugar común, es una idea tan aceptada que apenas merece la pena discutirla.

Una manera de ilustrar esta idea es buscando señales visibles que delaten esa transformación de fondo, como por ejemplo la evolución del perfil de las empresas más importantes del mundo, o el de las personas más ricas, y compararlo con el de hace 30 años.

Viendo la lista de las 10 personas más ricas del mundo (2), parece que hay un ámbito dominante claro: las tecnologías de la información.

TABLA 1
10 PERSONAS MÁS RICAS DEL MUNDO

Ranking	Nombre	País	Empresa	Actividad
1.	Jeff Bezos	USA	Amazon	Tecnología
2.	Elon Musk	USA	Tesla, SpaceX (anteriormente PayPal)	Tecnología y automoción
3.	Bernard Arnault & family	Francia	LVMH	Moda y retail (segmento lujo)
4.	Bill Gates	USA	Microsoft	Tecnología
5.	Mark Zuckerberg	USA	Facebook	Tecnología
6.	Warren Buffett	USA	Berkshire Hathaway	Finanzas e inversiones
7.	Larry Ellison	USA	Oracle	Tecnología
8.	Larry Page	USA	Google	Tecnología
9.	Sergey Brin	USA	Google	Tecnología
10.	Mukesh Ambani	India	Diversificado	Diversificado

Fuente: Forbes, 2021. https://www.forbes.com/billionaires/

TABLA 2
COMPAÑÍAS CON MAYOR CAPITALIZACIÓN DE MERCADO EN EL MUNDO

Ranking	Nombre	Capitalización (en trillones de dólares americanos, billones europeos)	
1	Microsoft	2,549 T$	USA
2	Apple	2,477 T$	USA
3	Saudi Aramco	1,977 T$	S. Arabia
4	Alphabet (Google)	1,972 T$	USA
5	Amazon	1,795 T$	USA

Fuente: https://companiesmarketcap.com/, a 17 de noviembre de 2021

En el número 11 está por cierto Amancio Ortega, que es de las pocas excepciones a la norma. La lista está claramente dominada por negocios basados en el intangible (independientemente de que todos requieren de infraestructuras y personas). La actividad de las empresas tiende a generar barreras de entrada para los nuevos competidores. Normalmente, cuando se produce la irrupción de nuevos actores se basa en algún tipo de innovación disruptiva que da lugar a un nuevo mercado todavía sin una organización dominante clara. Una vez se establece un nuevo líder, se levantan a su vez nuevas barreras de entrada para los siguientes (en otras palabras, es difícil que aparezca un nuevo facebook o un nuevo twitter, a menos que los que los desbanquen sean lo bastante diferentes a ellos). Por supuesto que la correlación no es causalidad, pero aún así parece muy significativo el perfil de estas personas cuya fortuna no proviene de una herencia familiar sino de la actividad de empresas que han creado ellos.

Las cinco compañías con mayor capitalización de mercado en el mundo (3) son Apple, Microsoft, Aramco, Alphabet y Amazon. Aramco es una corporación del sector de la energía, una petrolera de Arabia Saudí, la única «intrusa» en esta lista. Las otras cuatro están en la lista de sospechosos habituales cuando hablamos de explotación de datos, a la que podríamos añadir Meta (Facebook) o Netflix.

Si hacemos un listado por beneficios (4) en los últimos años, podemos observar que desde 2015 hay unos nuevos actores que nos resultan familiares.

Se aprecia que, poco a poco, empresas con uso intensivo del dato han ido entrando en este listado.

¿Hasta qué punto es una tendencia?

Si nos preguntamos por las compañías que están creciendo con mayor rapidez, el listado de la revista *Fortune* para 2021 es el siguiente, con algunos incrementos de beneficio mareantes para cualquier sector:

Veamos algunas empresas del listado quizá menos conocidas: Cigna es una multinacional del sector de la salud; StoneX se dedica a los servicios financieros, haciendo un uso extensivo de plataformas digitales; Progressive es una empresa del sector seguros (propiedad y accidente).

Las dos compañías de seguros que aparecen en la lista hacen un uso extensivo del BigData para sus operaciones (5).

La primera observación que podemos hacer es que casi todas ellas utilizan datos personales como parte de su modelo de negocio. La segunda es que no hay una sola empresa europea en este listado ni en los anteriores. En el ranking global 500 de Forbes, aparece Volkswagen (6) en la décima posición por ingresos (no por beneficio, que es la tabla que hemos visto).

Por beneficio es Nestlé la primera marca europea que entra en la lista, en la posición 29. Es un detalle importante, y parece claro que la Unión Europea también se ha dado cuenta y tiene interés en que el escenario cambie.

Así, otro modo de ver la importancia del dato en la economía actual es constatando su impacto en los planes estratégicos y la normativa, en este caso euro-

TABLA 3
COMPAÑÍAS CON MAYORES BENEFICIOS

Año 2015	Año 2018	Año 2021
Industrial & Commer. Bank of China	Apple	Apple
Apple	British American Tobacco	Saudi Aramco
China Construction Bank	Berkshire Hathaway	SoftBank Group
Exxon Mobil	Industrial & Commercial Bank of China	Industrial & Commercial Bank of China
Agricultural Bank of China	Samsung Electronics	Microsoft
Bank of China	China Construction Bank	Berkshire Hathaway
Wells Fargo	Verizon	Alphabet
Microsoft	AT&T	China Construction Bank
Samsung Electronics	Agricultural Bank of China	Agricultural Bank of China
JP Morgan Chase	Bank of China	Facebook

Fuente: Ranking Global 500, revista *Fortune*

TABLA 4
COMPAÑÍAS CON CRECIMIENTO MÁS RÁPIDO

Ranking	Empresa	Ingresos (MS)	Incremento de ingresos desde 2020	Beneficio (MS)	Incremento de beneficio desde 2020
3	Amazon	386.064	37,60 %	21.331	84,10 %
21	Alphabet	182.527	12,80 %	40.269	17,30 %
28	Cigna	160.401	4,50 %	8.458	65,70 %
63	Alibaba Group Holding	105.866	44,70 %	22.224	3,60 %
86	Facebook	85.965	21,60 %	29.146	57,70 %
190	StoneX	54.140	64,60 %	170	99,30 %
275	Progressive	42.658	9,30 %	5.705	43,70 %
484	Netflix	24.996	24,00 %	2.761	47,90 %

Nota: Los incrementos se calculan respecto a 2021, es decir, Amazon tuvo 280.522 M$ de ingresos y unos beneficios de 11.588 M$ en 2020, según los datos de *Fortune*.
Fuente: Ranking Global 500, revista *Fortune*

pea. La UE trata de buscar la armonía entre la explotación comercial de los datos, y la protección de los derechos de las personas (7). Ética y beneficio, ¿una tensión ancestral? Sin ánimo de entrar en detalle en esta introducción, la UE contempla los datos en su estrategia no sólo como generadores de riqueza, también como elemento clave geoestratégico cuando habla por ejemplo del concepto de «soberanía digital». Estar en el mundo, ser relevante, implica controlar y explotar tus datos (al menos los tuyos). Este interés estratégico se vislumbra también en las legislaciones sobre protección de datos, interoperabilidad o la directiva relativa a los datos abiertos y la reutilización de la información del sector público.

A su vez las estrategias se concretan en planes y líneas de financiación. Por un lado las ordinarias (fondos estructurales, Horizonte Europa, Programa Europa Digital (8)), y por otro las extraordinarias que se han puesto en marcha como consecuencia de la pandemia (un claro ejemplo es el instrumento NextGenerationEU y su desarrollo en España mediante el Plan de Recuperación, Transformación y Resiliencia, que priman los proyectos para la transición digital y verde de la economía).

Como colofón y tercer indicador de esta tendencia en el uso de los datos, y yendo un paso más allá en la dirección de la explotación de datos provenientes de distintas fuentes, se podrían repasar algunas iniciativas recientes (entre otras) relacionadas con repositorios de datos compartidos.

Dataspace DEP + GaiaX

Gaia-X (9) es un proyecto europeo en el que representantes de la empresa, la política y la ciencia trabajan juntos para crear una infraestructura de datos federada y segura. La finalidad es que las empresas y los ciudadanos puedan cotejar y compartir datos, de manera que mantengan el control, teniendo poder de decisión sobre lo que ocurre con ellos y dónde se almacenan.

La arquitectura de Gaia-X es descentralizada. Gaia-X es el macrosistema resultante de interconectar una multitud de plataformas individuales que siguen un estándar común: el estándar Gaia-X. El objetivo es desarrollar una infraestructura de datos basada en los valores de apertura, transparencia y confianza. Se trata

no tanto de crear una nube sino un sistema en red que conecte a muchos proveedores de servicios en la nube.

Dataspace INE + AEAT + GISS + BDE + SEPE

En este sentido, el Instituto Nacional de Estadística (INE) (10), la Agencia Estatal de Administración Tributaria (AEAT), la Seguridad Social (SS), Servicio Público de Empleo Estatal (SEPE) y el Banco de España (BE), como titulares de una cantidad importante de información administrativa y de datos granulares recogidos con fines estadísticos, han acordado comenzar a trabajar conjuntamente en el diseño de un sistema colaborativo de datos al que los investigadores tendrán acceso, siempre que la información vaya a ser utilizada con fines científicos de interés público. Además, promoverán la incorporación futura de otras instituciones que sean titulares de registros administrativos u otras bases de datos granulares de interés para la investigación.

EOSC

La European Open Science Cloud (11) es un proyecto europeo que aspira a proporcionar a los investigadores, innovadores, empresas y ciudadanos europeos un entorno multidisciplinar federado y abierto en el que puedan publicar, encontrar y reutilizar datos, herramientas y servicios con fines de investigación, innovación y educación.

Este entorno funcionará en condiciones bien definidas para garantizar la confianza y salvaguardar el interés público.

La EOSC permite promover un cambio en las comunidades científicas y las infraestructuras de investigación destinado a conseguir:

- acceso garantizado a los resultados y recursos de información empleados para la investigación.
- gestión FAIR (Findability, Accessibility, Interoperability and Reusability) que permite una reutilización fiable de los datos de investigación y de todos los demás objetos digitales producidos a lo largo del ciclo de vida de la investigación (por ejemplo, métodos, software y publicaciones).
- reproducibilidad de los resultados de la actividad científica.

Repositorio COVID

La Infraestructura Europea para la Investigación Clínica (ECRIN) ha creado un repositorio europeo de metadatos con el objetivo de compartir información de utilidad para el lanzamiento de ensayos clínicos relacionados con la COVID-19 en el que participa la Plataforma Española de Ensayos Clínicos (SCREN), dependiente del Ministerio de Ciencia e Innovación.

Hemos visto la importancia que tiene el tratamiento de los datos para Europa desde el punto de vista de la soberanía y de la necesidad de extraer valor de ellos tanto para generar riqueza como para ofrecer servicios de calidad. Cabe reflexionar sobre lo que podríamos llamar el «hecho diferencial europeo» en relación con esta actividad, en comparación con el planteamiento que tiene lugar en EEUU o China, por citar los dos referentes más importantes en la actualidad.

En estos países se realiza una explotación desde grandes plataformas que obtienen los datos a partir de servicios. Estas plataformas monetizan estos datos después fundamentalmente de dos maneras. En primer lugar, incorporando la información a sus cadenas de valor en el caso del tratamiento de datos de negocio. Por ejemplo empresas de seguros o fabricantes de automóvil los utilizan para mejorar sus productos o realizar ofertas a sus clientes. En segundo lugar vía actividades de marketing, usando los datos para el perfilado de los usuarios y la venta de publicidad personalizada (los anunciantes pagan por exponer su publicidad ante las personas con más probabilidad de adquirir sus productos en base a las determinaciones de los algoritmos).

El elemento que distingue el planteamiento europeo, y se puede decir con propiedad que Europa está liderando esta visión particular, es el respeto a la privacidad y el derecho de los ciudadanos a saber en todo momento quién tiene sus datos y para qué los está utilizando.

En este marco cobra especial sentido la creación de datasets impulsados desde la Administración con el fin de mejorar los servicios a los ciudadanos y optimizar los recursos públicos y la eficiencia del gasto, que es la manera de poder llegar a más personas.

La idea es que los datos trabajen más para las personas que los generan, no tanto (o únicamente) para las corporaciones. Está claro que las empresas europeas para ser competitivas tienen que incorporar el dato a su cadena de valor, no se trata de que se queden fuera. De hecho, las empresas tienen también mucho que aportar y mucho de lo que beneficiarse en relación con estos datasets. Pero pensemos ahora en lo que la inteligencia artificial y el uso del BigData pueden aportar a los ciudadanos aplicados al campo de la medicina, de la movilidad o de la justicia, la lucha contra enfermedades raras, mejora de la asistencia a los mayores y mejoras de la eficiencia de las infraestructuras (tratamiento de historial clínico, repositorio jurídico. Dado el nivel de avance tecnológico y el grado de digitalización de su administración, España dispone de los mayores repositorios en castellano en estos y otros ámbitos administrativos). Pero para ver resultados se requiere compartir datos respetando condicionantes como la privacidad o el secreto estadístico, con una gobernanza que involucra a distintas instituciones, lo que no siempre es sencillo. ¿Se puede conseguir utilizando repositorios centralizados?

PROBLEMAS DE LOS REPOSITORIOS O ENCLAVES CENTRALES DE DATOS

La experiencia actual permite constatar que es un enfoque que tiene sus limitaciones y aspectos críticos, relacionados fundamentalmente con la seguridad y la gobernanza.

La gestión eficaz y ética de estas dos cuestiones supone un esfuerzo añadido que en ocasiones limita el alcance y despliegue de este tipo de soluciones, bien por tiempos, bien por la necesidad de poner de acuerdo a actores muy diferentes, bien por los recursos de gestión administrativa, equipos y personas que requiere.

Desde el punto de vista de la seguridad los condicionantes son similares a los de la operación de cualquier tipo de repositorio, si bien hay que tener en cuenta que el compromiso de un sistema central puede poner en riesgo simultáneamente gran cantidad de conjuntos de información.

Hay que tener en cuenta que el número de puntos de carga y transmisión de datos es posiblemente mayor (en todo caso mayor que en el caso de los repositorios individuales que vuelcan en él) y además muy heterogéneo, en el sentido de que son agentes diferentes los encargados de la captación, custodia y transmisión del dato.

A esta circunstancia se añade el problema de la alta dimensionalidad (12) del repositorio centralizado, lo que permite que mediante el cruce de variables que no conviven en los diferentes repositorios que vuelcan en él, se puedan llevar a cabo ataques de reidentificación o desanonimización que no se podrían realizar sobre dichos repositorios por separado (13) , (14). Es decir, datos que se consideran «seguros» en un repositorio, en el sentido de convenientemente anonimizados, dejan de serlo al poderse cruzar con nuevos datos provenientes de otros repositorios.

No es sólo importante la cuestión de la seguridad, también es clave la actualización de esos datos. Es preciso encontrar las tasas de refresco óptimas y verificar colisiones cuando datos que se solapan entre repositorios tienen tasas de refresco distintas, lo que requiere diseñar una estrategia óptima para liberar los datos teniendo en consideración las tasas de refresco de todos los repositorios, resolver conflictos, etc.

Como se ha indicado, este es un problema inherente a los sistemas de intercambio de información y existen técnicas pensadas para atacarlo en función del tipo de datos (texto, audio, imagen, información geoespacial, datos genéticos) así como de su grado de estructuración (16).

En todos los casos, conseguir ese nivel de riesgo aceptable conlleva un esfuerzo, una inversión de recursos de algún tipo, y por ello resulta deseable que el nivel de sensibilidad de los datos que se intercambien sea lo más bajo posible, ya que el riesgo, y por tanto la inversión en seguridad, está directamente relacionado con la sensibilidad del dato a proteger. Esto resulta especialmente crítico cuando el tipo de datos que se quieren intercambiar es particularmente delicado, como sucede por ejemplo con los registros médicos y el resto de los considerados como «datos de especial protección» por el RGPD (17). Parece claro que no es lo mismo intercambiar registros CMBD (Conjunto Mínimo Básico de Datos) sanitarios que intercambiar el agregado de las operaciones estadísticas que se quiere realizar sobre ellos (por ejemplo, la incidencia de determinado tipo de cáncer en una región).

En este sentido, merece la pena recordar lo que el propio reglamento de protección de datos dice al respecto en el apartado 1 de su artículo 25 «Protección de datos desde el diseño y por defecto» (18), (19) (resalte del autor):

> *«Teniendo en cuenta el estado de la técnica, el coste de la aplicación y la naturaleza, ámbito, contexto y fines del tratamiento, así como los riesgos de diversa probabilidad y gravedad que entraña el tratamiento para los derechos y libertades de las personas físicas, el responsable del tratamiento aplicará, tanto en el momento de determinar los medios de tratamiento como en el momento del propio tratamiento, medidas técnicas y organizativas apropiadas, como la seudonimización, concebidas para aplicar de forma efectiva los principios de protección de datos, como la minimización de datos, e integrar las garantías necesarias en el tratamiento, a fin de cumplir los requisitos del presente Reglamento y proteger los derechos de los interesados».*

Esta referencia nos conduce de lleno al terreno de la gestión de dos aspectos no estrictamente técnicos, a saber, cumplimiento normativo, que hace de puente entre los dos mundos dado que tiene implicaciones técnicas, y la gestión administrativa, que igualmente tiene implicaciones técnicas.

En la Unión Europea existe una legislación de tipo garantista que pone el foco en los derechos de los ciudadanos, como demuestra el hecho de que la primera consideración que hace el RGPD es la siguiente «La protección de las personas físicas en relación con el tratamiento de datos personales es un derecho fundamental», lo que en absoluto resulta una afirmación baladí. No es este el espacio para considerar todas las implicaciones del RGPD y su transposición a la normativa española; Es interesante mencionar también la propuesta para la Ley de Gobernanza de Datos (20).

La gestión de los datos en sí, como consecuencia también de la normativa, conlleva tener en consideración los aspectos de gobernanza de los mismos. El uso transparente, «lícito y leal» al que se refiere el Reglamento, la obligación de informar, recoger el consentimiento (que «debe otorgarse mediante un acto afirmativo claro que refleje una manifestación de voluntad libre, específica, informada, e inequívoca del interesado»), y la garantía de los derechos de acceso, rectificación etc. (21), obligan a establecer sistemas de

trazabilidad y auditoría de acceso y uso de la información que habría que desplegar no sólo en los repositorios propiedad de los responsables de tratamiento donde residen los datos originales, sino también en el repositorio centralizado donde fueran volcados.

Una cuestión adicional es que los problemas de gobernanza, como ponían de manifiesto los participantes en el debate organizado por el Observatorio de Sector Público de Inetum en febrero de 2021, radica en que los datos son responsabilidad de cada uno de los organismos en los que se han depositado, «lo que dificulta tener una visión global y realizar actuaciones conjuntas para *sacar provecho y dar uso real a esos datos*» (cosa que se complica todavía más en el caso de que existan competencias concurrentes entre distintas administraciones) (22).

Por último, estos requisitos demandan de las organizaciones responsables de los tratamientos que van a compartir datos el dotarse de una serie de salvaguardas con el fin de poder demostrar que han actuado con la debida diligencia en el cumplimiento de las obligaciones derivadas de la norma. Esto supone la necesidad de suscribir acuerdos en forma de convenios, contratos u otro tipo de actos jurídicos vinculantes, lo que viene a añadir una capa adicional de complejidad en la gestión, necesidad de especificar a priori los análisis de información a realizar y el impacto que todo ello tiene en la dilatación de los tiempos de ejecución de los proyectos de análisis de información.

Todas estas consideraciones se traducen en la práctica en la constatación de que al final estos modelos sólo se han puesto en marcha para colecciones reducidas de datos, con frecuencias de actualización bajas (CMBD del Ministerio de Sanidad por ejemplo).

Recapitulando: hemos visto que los sistemas centralizados adolecen de algunos inconvenientes que se pueden resumir en:

- seguridad: un sistema centralizado requiere un enorme esfuerzo de protección de la información ya que si el sistema es comprometido se pueden ver expuestos todos los datos que lo componen.
- problemas de dimensión territorial y competencial, cuando requerimos compartir información de distintas administraciones. Exige la autorización administrativa de los ciudadanos participantes así como la firma de convenios o protocolos entre las administraciones participantes,
- en el caso de organizaciones que compiten en el mercado existe un problema de confianza en cuanto a depositar datos críticos de su negocio o referidos a sus clientes y compartirlos con los rivales.
- necesidad de consenso cuando el número de participantes sea muy alto, por ejemplo, cuando se trata la información de ciudadanos o usuarios de servicios de telefonía.
- granularidad del dato: es técnicamente complicado realizar de forma centralizada ciertas operaciones en tiempo real (como en grandes sistemas de control y adquisición de datos, por ejemplo *smart cities*) sobre grandes volúmenes de información producida por la sensórica o la actividad de un sistema distribuido (ej. red de telefonía o sistema de suministro eléctrico).

¿Podría haber un enfoque que resolviera o simplificara algunas de estas cuestiones, permitiendo realizar operaciones de explotación sobre repositorios de datos federados en lugar de sobre repositorios centralizados?

Pensamos que sí, que una solución podría venir precisamente del uso de espacios de datos federados que, con modelos de cálculo distribuido, control del secreto estadístico y el uso de nuevas técnicas criptográficas resolviera estos problemas.

ASPECTOS TÉCNICOS ESPECÍFICOS DE LOS ESPACIOS DE DATOS DISTRIBUIDOS

En relación con los espacios de datos distribuidos es necesario tener en consideración algunos aspectos específicos de este tipo de esquemas, como son:

- Definición de estructuras de información federadas
- Modelos de distribución de la información
- Seguridad en los sistemas de información distribuidos
- Uso de criptografía homomórfica y diferencial en los sistemas de información distribuidos
- Control de la confidencialidad estadística en sistemas de información distribuidos
- Aprendizaje máquina en los sistemas de información federados

Definición de estructuras de información distribuida

Cuando pretendemos utilizar diversos sistemas de información federados, lógica y geográficamente, es necesario realizar una buena gobernanza de los metadatos.

El Esquema Nacional de Interoperabilidad (23) (ENI) establece los principios y directrices de interoperabilidad en el intercambio y conservación de la información electrónica por parte de Administraciones Públicas.

Para poder operar con estas fuentes de información distribuidas necesitamos establecer una interoperabilidad a nivel de registro que incluya entre otras las siguientes normalizaciones:

- Coordinación de los metadatos y los tipos de datos a nivel de registro.
- Interoperabilidad de las clasificaciones empleadas y las codificaciones asociadas.

- En el caso de emplearse ontologías, estas deben a su vez ser coordinadas y emplear ontologías base coherentes.
- En caso de integrar la información de diversas fuentes, se debe coordinar el sistema de identificadores de registro. Teniendo en cuenta, si así es requerido por la protección de los datos personales, la seudoanonimización coordinada de tales identificadores y la perturbación de microdatos.

Esta información debe ser consolidada en catálogos de datos desde los que compartir los diseños de registros.

Algunas de estas ideas han sido plasmadas en estándares de coordinación de recursos de información federada como SDMX (24), empleado por Eurostat, así como algunos organismos nacionales (INE). La iniciativa SDMX establece normas técnicas (incluido el modelo de información), una arquitectura y herramientas informáticas para facilitar el intercambio de datos y metadatos estadísticos. Desde 2004 se han publicado varias versiones de las especificaciones técnicas. SDMX también se ha publicado como norma internacional ISO (ISO 17369 (25)).

Modelos de distribución de la información en la federación de datos

En función de su estructura o relación jerárquica se pueden considerar los siguientes tipos de modelos:

- Registros con estructura similar de información distribuida geográficamente (federación vertical). Se trata de registros de información con una estructura similar pero distribuidas geográficamente. Un ejemplo paradigmático es la información perteneciente a registros administrativos de ciudadanos y empresas que se hallan dispersos en organizaciones locales y autonómicas de la Administración.
- Registros con estructura de información distribuida complementaria (federación horizontal). Se trata de registros distribuidos que contienen información complementaria a nivel de registro. Un conjunto de información es el constituido por los diversos registros administrativos de cada ciudadano mantenidos por cada una de las organizaciones de la Administración General del Estado en el marco de sus competencias (ej. AEAT, Seguridad Social).
- Registros organizados jerárquicamente (federación jerárquica). En ocasiones existen diferentes niveles de información distribuida entre distintas organizaciones dentro de un mismo marco competencial. Un ejemplo sería la información detallada contenida en la historia clínica electrónica de la que se extrae un sustrato resumido (denominado Conjunto Mínimo Básico de Datos, CMBD (26)) reportado a las Consejerías de Salud, desde hospitales y servicios de atención primaria, así como al Ministerio de Salud y a terceros organismos internacionales. La característica peculiar de este tipo de distribuciones se encuentra en la dependencia jerárquica de la información y su localización distribuida.
- Datos contenidos en un sistema de fedatario distribuido (o de cadena de bloques). Una fuente de información distribuida, que comienza sus andaduras en la Administración, es la contenida en sistemas de fedatario distribuido, implementados por medio de técnicas de cadenas de bloques. Tanto si se trata de un sistema permisionado como no permisionado, este tipo de sistemas ofrecen una nueva forma de federación de registros no centralizada físicamente que vienen a reimplementar los sistemas transaccionales tradicionales y nuevas formas de compromiso público entre las partes (ej. *Smart Contracts*).

Las tres primeras formas de distribución de información son operables con las técnicas hoy en día existentes. Por una parte, los principales vendedores de sistemas de OLAP (*Online Analytical Processing*, Procesamiento Analítico en Línea) y BI (*Business Intelligence*, Inteligencia de Negocio) admiten la hibridación de conjuntos multidimensionales o cubos de forma distribuida. Las consultas sobre estas bases de información son descompuestas, enviadas a las respectivas fuentes (previamente definidas en dichos cubos híbridos), agregadas y mostradas al usuario. Por otra parte, los principales lenguajes usados por desarrolladores e investigadores para la depuración y explotación de datos estructurados (ej. Python) disponen de librerías de procesamiento de datos estructurados distribuidos (ej. Dask, Spark (27)).

Respecto a la última forma de distribución de datos, su explotación distribuida depende de la estrategia de almacenamiento, encriptación de los contenidos y distribución del libro mayor digital (lista de registros electrónicos que constituye la cadena de bloques).

Seguridad en los sistemas de información distribuidos

De forma coordinada con la definición de la interoperabilidad, a nivel de registro, y el modelo de distribución de datos federado, deben establecerse con claridad las reglas de control de acceso y auditoría del sistema de información federado. En el caso de sistemas de información físicamente distribuidos, la seguridad es un factor todavía más crítico. La traza de actividad y el control de acceso registrado por cada uno de los nodos participantes en el sistema debe ser coherente y globalmente coordinado.

Existen diversas alternativas técnicas para solventar el problema de distribución del sistema de autenticación y autorización; sistemas criptográficos simétricos (ej. Kerberos) así como otros protocolos basados en el intercambio asimétrico de claves —ej. SSL, encrypted key exchange (EKE), secure remote password (SRP)—. La reciente incorporación de métodos multifactor, también llamados MFA (en ocasiones simplemente se usa 2FA, *two-factor authentication*) puede mejorar la

seguridad del sistema permitiendo presentar múltiples evidencias de las credenciales de los usuarios del sistema federado.

Por otra parte, la traza de la actividad de los usuarios en el sistema federado debe nutrirse de la información aportada por cada uno de los subsistemas de traza o log individuales que componen el sistema de información federada. En ocasiones, además de la traza de actividad del sistema por parte de los usuarios, se recoge la actividad de los servicios incluyendo alertas y errores del mismo. En este punto, la visión del sistema debe ser global dado que se emplea como un todo.

La localización del almacenamiento de las trazas de actividad (o logs) del sistema depende del volumen de dicha información. Existen sistemas de colas de datos centralizadas escalables (ej. Kafka (28)) o modelos de distribución de las trazas de forma local a cada uno de los nodos del sistema federado (ej. Apache BookKeeper (29)). En cualquiera de los dos modelos es necesario definir estándares para coordinar el almacenamiento de la actividad desarrollada por el usuario. Los requisitos de traza y control de acceso vienen parcialmente marcados, entre otros, por los requerimientos establecidos por la protección de datos personales.

Criptografía homomórfica y diferencial en los sistemas de información distribuidos

Uno de los problemas tradicionales de los sistemas centrales de información, cuando existen múltiples fuentes de información distribuidas, se encuentra en la necesidad de que los participantes en el sistema depositen en un nodo central todos los datos. Ello requiere una confianza absoluta en el control del acceso a la información y la seguridad del sistema central. En el caso de la Administración, adicionalmente, es necesaria la firma de convenios o protocolos administrativos para el tratamiento de la información recabada con fines específicos distintos a los inicialmente establecidos.

Tanto la criptografía homomórfica (30) como la diferencial (31) permiten realizar operaciones sobre los datos cifrados y obtener resultados, también cifrados, *equivalentes a los que se obtendrían de haber realizado las operaciones sobre los datos sin cifrar*. Estos resultados pueden ser posteriormente desencriptados por el propietario de las claves para acceder a la información resultante, sin que otros participantes en el sistema hayan tenido acceso a ella a lo largo del proceso. Conseguimos por tanto un espacio de operación encriptado donde se conjuga el acceso a los datos y se salvaguarda la información personal.

Cuando la funcionalidad buscada en la explotación del sistema de información distribuido es la propia de un gestor de base de datos, por ejemplo para la consulta de registros agregados, existe la posibilidad de emplear bases de datos encriptadas (ej. CryptDB (32), Google Encrypted BigQuery (33), MS SQL Server Always Encrypted (34)).

Para la realización de estadísticas descriptivas, donde las operaciones son esencialmente cruce de identificadores y agregación de datos (incluyendo porcentajes, medias, etc.), se propone el uso de este esquema de encriptación para garantizar la protección de datos personales y posibles compromisos de la seguridad informática del sistema. En este caso, se prefiere el uso de criptografía parcialmente homomórfica (PHE (35)), cuyo coste computacional es mucho más reducido que la criptografía completamente homomórfica (FHE (36)), u otras técnicas de Privacy-enhancing technologies (PET). Estos esquemas no se hallan integrados con la mayor parte de los motores OLAP pero es posible el desarrollo de scripts que implementen estas técnicas y realicen operaciones de forma programática.

En el caso de operaciones estadísticas de mayor complejidad será necesaria la utilización de criptografía completamente homomórfica (FHE) sobre aproximaciones polinómicas. El uso de este tipo de criptografía requiere en la práctica de la aproximación polinómica de funciones así como de fuertes capacidades computacionales. Existen implementaciones en forma de librerías de uso abierto (ej. SEAL, Palisade, HELib, Pyfhel (37)).

Respecto a la integración de este tipo de técnicas criptográficas en los sistemas de inteligencia de negocio, como se ha comentado, existe la posibilidad de integración en los propios motores de base de datos (por ejemplo, empleando los populares motores de MS o MySQL (38) (39)) o en los propios sistemas de inteligencia de negocio de forma nativa (ej. Google Cloud BigQuery (40)).

En lo relativo al uso de este tipo de técnicas en el ámbito del aprendizaje automático, existen algunos sistemas ya entrenados sobre criptografía homomórfica (41). Entre los paradigmas implementados se encuentran sistemas basados en regresión logística, árboles de decisión o redes convolucionales. Estos sistemas permiten encriptar los datos en origen y entrenar un modelo en un sistema remoto encriptado. El modelo obtenido puede operar, local o remotamente (por ejemplo, desde la nube) sobre datos encriptados. La principal contrapartida de este tipo de sistemas es la fuerte carga computacional. Las operaciones no lineales deben implementarse con FHE.

Control de la confidencialidad estadística en sistemas de información distribuidos

El control de la confidencialidad es una tarea necesaria en la gestión de datos que requieran protección de datos personales o económicos de las empresas contenidos en los registros individuales o pequeñas agregaciones. Hay que alcanzar un compromiso entre la creciente demanda de información económica y social, por medio de una amplia y precisa información estadística, y la obligación legal y ética de proteger la privacidad de los individuos y las empresas que son la fuente de dichos datos estadísticos. No se trata, por tanto, de un problema de confidencialidad exclusiva-

mente, sino de maximizar la información proporcionada manteniendo las restricciones que impone la protección de los datos estadísticos confidenciales.

En el ámbito legal, el control de la confidencialidad estadística en la difusión de datos estadísticos está regulado, a nivel estatal, por medio de la Ley de la Función Estadística Pública 12/1989 (42), Título 1, Capítulo III «Del secreto estadístico», y en el ámbito europeo es el Reglamento 233/2009 del Parlamento Europeo (43), en su capítulo V, artículo 20.

El problema del control de la confidencialidad estadística consiste en la no divulgación de datos confidenciales directos así como de datos que permitan la inferencia de datos confidenciales. En lo relativo a las técnicas de control de la confidencialidad estadística, el principal objetivo es aportar la máxima protección de los datos confidenciales con la menor pérdida de información.

El conjunto de métodos que intentan dar solución a este problema se denominan técnicas de control de la confidencialidad estadística (Statistical Disclosure Control, SDC). En la actualidad existe una gran variedad de técnicas de control de la confidencialidad estadística (44):

- Métodos de Restricción: este tipo de métodos consiste en la limitación de la información publicada. Podemos subdividirlos a su vez en:
    - Métodos de recodificación: agrupación de valores de variables de clasificación para ocultar los datos confidenciales de determinados colectivos minoritarios, incluyendo la eliminación de algunos valores de la variable o variables elegidas.
    - Supresión de celdas (CSP): se suprimen tanto las celdas que contienen información confidencial (supresión primaria), como aquellas que puedan servir para inferir las primeras (supresiones secundarias).
    - Publicación de intervalos: en lugar de publicar datos individuales, se muestran los márgenes del intervalo seguro en el que se encuentra el verdadero valor.
- Métodos de perturbación: los métodos de control de la confidencialidad estadística basados en la perturbación de datos emplean la alteración de los datos originales proporcionando unos nuevos datos sintéticos protegidos. Dentro de este conjunto de métodos se pueden citar los siguientes:
    - Redondeo aleatorio (CRP): se trata de redondear de forma aleatoria los valores de los datos confidenciales para mantener su confidencialidad.
    - Ruido aleatorio (CTA): emplea la introducción de pequeñas variaciones en los valores o en las variables de clasificación de modo que desaparezcan o queden ocultas las celdas deseadas.
    - Sustitución: intercambio de valores dentro de una variable de clasificación para su tratamiento estadístico sin posibilitar su identificación.
    - Microagregación: consiste en sustituir los valores confidenciales por la media de los valores dentro del grupo al que pertenecen.

En cuanto a la implementación de estos métodos existen diversas implementaciones abiertas entre las que cabe citar:

- TauArgus (45) (46): software diseñado para proteger tablas estadísticas.
- JCSP (47): aplicación para el control de la confidencialidad estadística en datos tabulares empleando la técnica de supresión de celdas. Basado en el optimizador lineal abierto GLPK.
- MuArgus (48): software diseñado para crear archivos de microdatos seguros.

Hasta donde sabemos, no existe una integración con los principales sistemas OLAP actuales. En el caso de JCSP, el sistema ha sido integrado con éxito con Pentaho (49), además de emplear técnicas de cálculo distribuido en cluster basadas en MPI (50). Se desconoce la existencia de aplicativos especialmente diseñados para entornos de información federada. Existe la posibilidad de adaptar los sistemas actuales al control de la confidencialidad estadística en un nodo central de confianza (requiere la identificación unívoca, en un esquema común de información, de cada registro sobre el que quiere garantizarse la confidencialidad estadística).

Además de estos retos técnicos, existe un reto organizativo en lo referente a la confidencialidad estadística y la gobernanza de los datos. La confidencialidad estadística está relacionada con los datos agregados que se presentan pero también con el conocimiento externo de otras fuentes de información que pueda tener un usuario atacante. El sistema de gobernanza del dato propuesto debe proporcionar una infraestructura de normalización a nivel de registro que aúne las estructuras de información existentes y permita identificar cada dato numérico conociendo en qué agregaciones participa. El sistema de información distribuido debe registrar de forma central el catálogo de datos publicados, pudiendo rastrear las operaciones en las que participa cada microdato, para poder realizar un cálculo seguro de la confidencialidad estadística.

Aprendizaje máquina en los sistemas de información distribuidos

Existe una gran cantidad de métodos para realizar el proceso de aprendizaje e inferencia sobre conjuntos de datos federados entre distintas localizaciones. En la

literatura, este tipo de aprendizaje puede encontrarse entre las publicaciones bajo el epígrafe de aprendizaje federado (51) o colaborativo. La distinción principal respecto a otros métodos de aprendizaje distribuido se encuentra en la alta latencia de comunicación, en comparación con un sistema paralelo tradicional (incluyendo clústeres de baja latencia), y la imposibilidad de comunicar entre los nodos federados la información relativa a los ejemplos empleados para el entrenamiento o las tareas de inferencia.

Un buen ejemplo de este tipo de entorno puede ser un conjunto de hospitales entrenando y aplicando modelos de aprendizaje automático sobre la historia clínica de sus respectivos pacientes. En este caso, no se desea colocar en un repositorio central toda la información contenida en la historia clínica electrónica (HCE) ni tampoco dar acceso abierto a investigadores y empresas interesados en desarrollar sus modelos de aprendizaje automático. Esta información se encuentra especialmente protegida en Europa, por el Reglamento general de protección de datos (RGPD), así como en EEUU, por el Health Insurance Portability and Accountability Act of 1996 (HIPAA).

El aprendizaje federado permite a múltiples actores construir un modelo de aprendizaje automático común sin compartir datos, sin comprometer la privacidad de los datos y garantizando el derecho de acceso a los mismos. Existen diversas tipologías de aprendizaje federado dependiendo, entre otros, de factores como la existencia de nodos centrales, el grado de heterogeneidad de los datos o el uso de técnicas criptográficas. A continuación se enuncian algunas de las principales tipologías (52) (53):

- Aprendizaje federado centralizado: se utiliza un servidor central para orquestar la computación necesaria para calcular el modelo. La selección de los nodos participantes así como la actualización de los parámetros del modelo son responsabilidad de un servidor central. Este tipo de sistemas de aprendizaje federado tiene una alta dependencia de la latencia de la red. Por otra parte, bajo este paradigma es posible ejecutar la mayor parte de los modelos de aprendizaje automático.

- Aprendizaje federado descentralizado: en este tipo de sistemas, un conjunto de nodos se coordina para calcular un modelo neuronal único. Las actualizaciones del modelo se intercambian sólo entre los nodos directamente, sin emplear un servidor central. En la actualidad este tipo de modelos de aprendizaje federado también se está empleando sobre conjuntos de datos distribuidos mediante cadena de bloques (*blockchain*) (54) (55).

- Aprendizaje federado heterogéneo: este tipo de sistemas de aprendizaje automático han sido diseñados para entornos de información heterogénea, por ejemplo, los teléfonos móviles. Entre los retos a los que se enfrentan este tipo de sistemas se encuentra la heterogeneidad de los datos, las distintas capacidades de cálculo de cada nodo así como los recursos de comunicación (56).

El aprendizaje federado entre silos de información o colecciones masivas de dispositivos (ej. red de telefonía o red de suministro eléctrico) presenta grandes desafíos prácticos y algorítmicos (57). Los modelos de aprendizaje van desde los modelos distribuidos totalmente descentralizados a aquellos diseñados entre pares iguales. La heterogeneidad de los conjuntos de datos supone también un reto para los algoritmos de aprendizaje automático. Los métodos de optimización del aprendizaje federado son muy diversos y sus rendimientos varían dependiendo de la distribución de los ejemplos de entrenamiento, la latencia de las comunicaciones, la compresión de los modelos y la heterogeneidad de los datos de entrenamiento.

Por último, pero no menos importante, la preservación de la privacidad de los datos del usuario requiere de un entorno en el que los cálculos se realicen de forma segura. En el caso de sistemas federados de aprendizaje es especialmente crítica la protección contra actores externos malintencionados o ataques basados en privacidad diferencial central. Así mismo, dada la heterogeneidad de los datos, deben tomarse medidas para garantizar la equidad e inexistencia de sesgo en los modelos calculados.

CONCLUSIONES Y PROPUESTAS EN RELACIÓN CON LAS PRÓXIMAS INICIATIVAS NACIONALES Y EUROPEAS DE TRATAMIENTO DE DATOS

En la actualidad, en España, existe una oportunidad única para avanzar en la creación y operación de espacios de datos federados.

Por una parte, la UE se esfuerza por mejorar drásticamente sus capacidades digitales, tanto en el despliegue de tecnologías digitales como en las competencias digitales necesarias para toda la mano de obra de la UE. Por ejemplo, en el recientemente publicado programa de trabajo Digital Europe Programme (58), así como en Connecting Europe Facility 2021-2027, (59), se promueve el desarrollo de infraestructuras digitales clave para innovar y reforzar su base industrial. Se trata de una enorme inversión público-privada que ningún Estado miembro podría acometer por sí solo. En este contexto, dentro de la Estrategia Europea de Datos (60) se ha anunciado un proyecto de alto impacto sobre los espacios de datos europeos, también llamados dataspaces, que abarca las arquitecturas de intercambio de datos y los mecanismos de gobernanza, así como la federación europea de infraestructuras de nube, eficientes energéticamente y seguras.

Por otra parte, a medio plazo la UE, por medio del instrumento NextGenerationEU (61), está impulsando la recuperación con el mayor paquete de estímulo jamás financiado en Europa. Se trata de un total de 2,018 billones de euros para reconstruir la Europa posterior a la COVID-19. Entre sus pilares se incluye el desarrollo de una Europa más ecológica, digital y más resiliente.

Finalmente, en el ámbito español, este impulso se ha traducido en la elaboración del Plan de Recuperación, Transformación y Resiliencia (62). En el componente C12, se desarrollan los espacios de datos sectoriales. Se trata de una contribución a los proyectos tractores de digitalización de los sectores productivos estratégicos. Este componente se centra en la puesta en marcha de grandes espacios de datos comunes orientados a la industria y seguros, que ayuden al impulso de la innovación empresarial en los sectores productivos estratégicos de la economía, entre ellos, el sector agroalimentario, el sector de la movilidad sostenible, el sector salud y el sector turístico. Se pretende asegurar la economía de escala y la eficiencia en la inversión mediante el desarrollo y compartición de componentes comunes (*building blocks*) y de infraestructuras y datos comunes, como las plataformas de servicios en la nube y las redes de supercomputación y almacenamiento, análisis de datos masivo y servicios basados en inteligencia artificial.

Es razonable esperar que el desarrollo de estos espacios de datos siga dos paradigmas paralelamente. Por una parte, la creación de enclaves computacionales centralizados y por otra la creación de espacios de datos federados.

El impulso de este segundo tipo de espacios de datos requerirá, entre otros, del desarrollo de los siguientes elementos:

Infraestructura técnica

Como se ha explicado en el apartado 3, es necesaria la definición y gobernanza de estructuras de información federadas. Más allá del Esquema Nacional de Interoperabilidad, que establece los principios y directrices de interoperabilidad, debemos impulsar también una interoperabilidad a nivel de registro. La creación de una infraestructura de información geográficamente distribuida (por ejemplo, que incluya diferentes organizaciones; central, local y autonómica) requiere la coordinación de las estructuras de registros de información distribuida, estructuras de información organizadas jerárquicamente así como los sistemas de reciente desarrollo basados en mecanismos de fedatario distribuido (o de cadena de bloques). Deben crearse esquemas y catálogos de información centrales con un gobierno eficaz del dato preparado para su procesamiento distribuido.

El espacio de datos federado debe proporcionar capacidades de explotación de la información desde la perspectiva tradicional OLAP o inteligencia de negocio. Para ello se pueden emplear herramientas de gestión de datos multidimensionales existentes en el mercado como los principales marcos de trabajo de programación distribuida para que los desarrolladores e investigadores puedan realizar tareas de depuración y explotación de datos estructurados.

Un sistema federado de información es un reto en lo relativo a la seguridad. Se debe vigilar la sincronización de las reglas y la estructura de los registros para la gestión del control de acceso y la auditoría de la actividad desarrollada en el sistema de información federado. La infraestructura técnica debe proporcionar capacidades de criptografía homomórfica y diferencial para el tratamiento distribuido de datos encriptados. Este tipo de sistemas ofrecen una capa adicional de control de la confidencialidad de los datos. El uso de este tipo de criptografía puede aligerar la carga administrativa producida por la firma de convenios o protocolos admistrativos para el tratamiento de la información.

El espacio de datos distribuido deberá proporcionar herramientas comunes para el control de la confidencialidad estadística (Statistical Disclosure Control, SDC). Entre los retos que se debe acometer se encuentra la integración con las herramientas OLAP actuales, la optimización de los tiempos de cálculo o la traza centralizada de los microdatos publicados de forma agregada.

En lo referente, al aprendizaje máquina se deben proporcionar capacidades de entrenamiento e inferencia sobre conjuntos de datos federados. Ha de trabajarse en la mitigación de los problemas de latencia y seguridad propios de los sistemas distribuidos.

Por tanto, se podría conceptualizar la propuesta como un sistema con tres capas diferenciadas:

1. una capa de datos, formada por los repositorios federados de distintos agentes que se desea integrar en el sistema
2. una capa técnica donde pueden construirse cubos y otros agregados de fuentes diversas y en la que operarían distintos subsistemas encargados de proveer de capacidad de análisis y servicios que operan sobre la información distribuida (ej. inteligencia de negocio, criptografía homomórfica, control de secreto estadístico (SDC), aprendizaje máquina federado (FML)).
3. una capa de gobernanza donde residen las herramientas de autorización, control de acceso monitorización, etc.

Un caso de uso de ejemplo puede ser el siguiente. Imaginemos una red de hospitales, perteneciente al Sistema Nacional de Salud, que quieren federar sus datos de diferentes servicios (ej. atención primaria, pruebas diagnósticas, informes de alta de urgencias) para el análisis de la evolución de distintas enfermedades.

1. Del repositorio completo de datos de cada hospital, se seleccionaría la parte que se va a federar en el sistema.
2. Se acuerdan las estructuras de información distribuidas y las normas de interoperabilidad definidas.
3. Se definen y emplean las técnicas de análisis pertinentes para resolver cada problema (BI, aprendizaje máquina y criptografía).

4. La capa de gobernanza se ocupa de que los grupos de investigación y empresas que van a estudiar el problema tengan acceso a los datos cifrados.

5. La red hospitalaria tendría acceso a los resultados cifrados de las operaciones de análisis sobre los datos, y la capacidad de descifrar dicho resultado, ya que es el único propietario de las claves.

Es muy importante resaltar que el sistema federado completo es mucho más que la suma de las partes. Pongamos como ejemplo paradigmático el estudio de las enfermedades raras. Por su propia naturaleza, ningún repositorio por separado tiene un número de datos elevado, y sabemos que para el aprendizaje máquina, por ejemplo, es necesario un gran número de casos para el entrenamiento de los algoritmos. El sistema propuesto permitiría poder realizar el entrenamiento utilizando los datos federados aportados por un gran número de hospitales.

Instrumentación administrativa y gobernanza

El recientemente nombrado Data Chief Officer de la Administración, o DCO (63), puede ser una figura relevante para la creación e impulso de los espacios de datos federados. Estos desarrollos deben ser coordinados con otros organismos internacionales, por ejemplo por medio de las normas internacionales ISO, con organismos europeos, como Eurostat o ETSI, así como con aquellos organismos nacionales dentro del marco de sus competencias (ej. INE). Este proceso de normalización e interoperabilidad puede ser cogobernado por los organismos competentes y coordinado por el DCO.

Dependiendo del dominio de la información contenida en el espacio de datos, la competencia administrativa para la gestión de los datos correspondientes a empresas y ciudadanos corresponderá a entidades locales, comunidades autónomas o la Administración General del Estado. En algunos de los sectores estratégicos, señalados por el Plan de Recuperación, Transformación y Resiliencia, en su componente C12. I1, como son la Sanidad y el Turismo, nos encontramos con una descentralización completa de las competencias. Los organismos centrales como ministerios u organismos de investigación deben jugar un papel importante en la coordinación de los desarrollos, el impulso I+D y la gobernanza de los recursos gestionados por estos sistemas. Por otra parte, tal y como se ha dicho, existen en la actualidad actuaciones paralelas en la UE. La interlocución con organismos europeos, tanto los competentes en dichos sectores estratégicos como los responsables del desarrollo de las infraestructuras digitales, debe ser coordinada por la Administración General (SEDIA y Ministerios en el marco de sus competencias)..

Sería aconsejable la formación de un comité de expertos, que respaldara las decisiones técnicas para el desarrollo de esta infraestructura, así como un Comité Directivo en el que tuvieran representación los tres niveles administrativos y las principales asociaciones empresariales e industriales que conforman cada uno de los sectores estratégicos.

Modelo de negocio y sostenibilidad

Será el desarrollo de modelos de negocio consistentes el que permita garantizar la sostenibilidad futura del sistema distribuido. El objetivo último es promover la innovación y reforzar la base industrial en este sector del país. Como se ha dicho, el esfuerzo debe fundamentarse en este tipo de arquitecturas de intercambio de datos, nuevos mecanismos de gobernanza, así como la federación de infraestructuras de nube energéticamente eficientes y seguras en el territorio europeo. Para que todo ello se sostenga en el tiempo obteniendo una verdadera soberanía del dato debemos desarrollar nuevos modelos de negocio. Este elemento es clave para poder incorporar a pymes y emprendedores en esta nueva economía del dato y sacar partido desde la singularidad europea; su grado de desarrollo del concepto de ciudadanía y el valor de los servicios de los que se le ha dotado, proporcionados por la Administración Pública.

La Administración puede promover una infraestructura de información distribuida tecnológicamente neutra e independiente de los intereses particulares de mercado, velando por el mantenimiento de la competencia y mitigando los riesgos de la misma. La Administración, por medio de sus órganos propios, puede ofrecer servicios transversales que apoyen tareas comunes, entre otras, como la asesoría en el ámbito de la protección de datos personales, la interoperabilidad, la normalización, los aspectos éticos o la seguridad.

Modelo de puesta en marcha

Tal y como acabamos de ver, la posibilidad de montar un sistema que permita realizar explotación de datos (desde operaciones estadísticas hasta aprendizaje máquina) sobre un modelo federado, respetando la privacidad de los datos, es muy real.

La clave es crear desarrollar un modelo de negocio sostenible que permita que el sistema se mantenga, una vez dado el impulso inicial de la Administración y de la UE.

¿En qué consistiría, por tanto, el rol de una agencia u organismo público en la puesta en marcha de este sistema?

En primer lugar, aportaría la confianza necesaria para atraer a los primeros agentes, especialmente en el caso de las entidades privadas. En el caso de las públicas, aportaría también el acceso a repositorios de datos.

Al ser una iniciativa promovida desde la Administración sobre la base del interés público, no existe conflicto de intereses ni competencial. El interés público en este caso podría enunciarse como «promover ciertas tec-

nologías para el mejor aprovechamiento y explotación de los datos con el fin de mejorar los servicios a los ciudadanos así como promover el desarrollo económico y el bienestar social».

En segundo lugar, actuaría como elemento neutro del sistema, precisamente por los mismos motivos. Sería el responsable de garantizar la interoperabilidad unificando protocolos, modelos, plantillas, estándares, metadatos, taxonomías... todos los elementos descritos en este artículo.

Por último, se podría plantear una cooperación público-privada en la que la Administración asumiera los costes iniciales de puesta en marcha de la infraestructura, con un retorno parcial (o no) de la inversión una vez alcanzado un punto de *break-even*. De este modo la Administración asumiría los costes iniciales con el objetivo de promover unas tecnologías en un campo estratégico como lo es la explotación de datos (64) y abriendo además un camino para un posterior desarrollo económico que también revertiría en la sociedad.

En resumen: la Administración tiene un rol que jugar como tractora de la innovación planteando y asumiendo retos y generando espacios de confianza donde resolverlos, en este caso la creación de espacios de datos distribuidos. Esto permite reducir riesgos y desarrollar nuevas soluciones con potencial, que luego la industria pueda convertir en bienes y servicios capaces de mejorar la calidad de vida de las personas. En el caso concreto de la explotación de datos, garantiza además el uso ético y los fines sociales, permitiendo la investigación y la mejora de los servicios. Pensamos que existe la necesidad, el interés y la oportunidad (especialmente considerando la disponibilidad de financiación a cargo de los Fondos de Recuperación y Resiliencia), de poner en marcha un sistema como el propuesto en este artículo que permita la explotación de datos cumpliendo las salvaguardas éticas, la protección de datos y el secreto estadístico, simplificando además la gestión administrativa.

NOTAS

[1] (Ackhoff 1989)
[2] Forbes, 2021. https://www.forbes.com/billionaires/
[3] https://companiesmarketcap.com/, a 17 de noviembre de 2021
[4] Ranking Global 500, revista *Fortune*
[5] https://builtin.com/big-data/big-data-insurance
[6] El grupo Volkswagen incluye las marcas Audi, Bentley, Bugatti, Ducati, Lamborghini, Porsche, SEAT y Škoda, y los camiones Scania y MAN.
[7] Una de las Seis Prioridades de la Comisión para 2019-2024 es «Una Europa Adaptada a la Era Digital», que se traduce, entre otras, en la estrategia europea de datos: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_es
[8] Programa Europa Digital : https://ec.europa.eu/info/funding-tenders/find-funding/eu-funding-programmes/digital-europe-programme_es
[9] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html
[10] Anuncio oficial del proyecto: https://www.bde.es/f/webbde/GAP/Secciones/SalaPrensa/NotasInformativas/21/presbe2021_32.pdf
[11] *(European Open Science Cloud (EOSC)*, s. f.), https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_es
[12] (Danezis *et al*. 2014)
[13] (Bindschaedler *et al*. 2018)
[14] (Al-Rubaie and Chang 2019)
[15] (Curzon, Almehmadi, and El-Khatib 2019)
[16] (Cunha, Mendes, and Vilela 2021)
[17] (Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos) (Texto pertinente a efectos del EEE), 2016)
[18] *(Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos) (Texto pertinente a efectos del EEE) 2016)*
[19] (Reglamento (UE) 2016/679, 2016), https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&-from=FR#d1e3126-1-1
[20] (EUR-Lex - 52020PC0767 - ES - EUR-Lex, 2020); https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A52020PC0767
[21] Derechos de: acceso, rectificación, supresión, oposición, derecho a la limitación del tratamiento y derecho a la portabilidad. Ver LEY ORGÁNICA 3/2018 DE PROTECCIÓN DE DATOS Y GARANTÍA DE DERECHOS DIGITALES (Jefatura del Estado 2018)
[22] (OSPI 2021)
[23] (BOE.es - BOE-A-2010-1331 Real Decreto 4/2010, de 8 de enero, por el que se regula el Esquema Nacional de Interoperabilidad en el ámbito de la Administración Electrónica., 2010); https://www.boe.es/buscar/act.php?id=BOE-A-2010-1331
[24] https://sdmx.org/
[25] https://www.iso.org/standard/52500.html
[26] *(Ministerio de Sanidad, Consumo y Bienestar Social - Portal Estadístico del SNS - Registro de Altas de los Hospitales del Sistema Nacional de Salud. CMBD, s. f.),* https://www.mscbs.gob.es/estadEstudios/estadisticas/cmbdhome.htm
[27] https://dask.org/
[28] https://kafka.apache.org/
[29] https://bookkeeper.apache.org/
[30] (aepd, 2020), https://www.aepd.es/es/prensa-y-comunicacion/blog/cifrado-privacidad-iii-cifrado-homomorfico
[31] https://en.wikipedia.org/wiki/Differential_privacy
[32] http://css.csail.mit.edu/cryptdb/
[33] https://cloud.google.com/bigquery/docs/encryption-at-rest

[34] https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-database-engine?redirectedfrom=MSDN&view=sql-server-ver15
[35] https://en.wikipedia.org/wiki/Homomorphic_encryption#Partially_homomorphic_cryptosystems
[36] https://en.wikipedia.org/wiki/Homomorphic_encryption#Fully_homomorphic_encryption
[37] https://github.com/microsoft/SEAL
[38] https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-database-engine?redirectedfrom=MSDN&view=sql-server-ver15
[39] (Popa *et al.*, 2011), https://people.csail.mit.edu/nickolai/papers/raluca-cryptdb-tr.pdf
[40] https://cloud.google.com/bigquery/docs/encryption-at-rest
[41] https://zhigang-chen.github.io/FHE%20and%20Machine%20Learning%20References.html
[42] (Ley 12/1989, de 9 de mayo, de la Función Estadística Pública, 1989), https://boe.es/buscar/doc.php?id=-BOE-A-1989-10767
[43] (Union, 2003), https://www.boe.es/doue/2009/087/L00164-00173.pdf
[44] (Willenborg & Waal, 1996)
[45] https://github.com/sdcTools/tauargus
[46] https://research.cbs.nl/casc/tau.htm
[47] (PAe - CTT - General - JCSP, s. f.), https://administracionelectronica.gob.es/ctt/jcsp#.YZyl0r3Mljg
[48] https://github.com/sdcTools/muargus
[49] (Porras Escalada *et al.*, 2012), https://www.ine.es/ss/Satellite?L=0&c=INERevEstad_C&p=1254735226759&cid=1259941931757&pagename=ProductosYServicios%2FPYSLayout#:~:text=Desarrollo%20de%20una%20aplicaci%C3%B3n%20de%20c%C3%B3digo%20abierto%20para%20el%20control%20de%20la%20confidencialidad%20estad%C3%ADstica%20en%20la%20difusi%C3%B3n%20de%20datos%20de-%20la%20AEAT
[50] Message Passing Interface, https://www.mcs.anl.gov/research/projects/mpi/; https://hpc-tutorials.llnl.gov/mpi/
[51] https://en.wikipedia.org/wiki/Federated_learning
[52] (Kairouz *et al.*, 2021), https://arxiv.org/pdf/1912.04977.pdf
[53] https://musketeer.eu/
[54] (Lo *et al.*, 2021), https://arxiv.org/abs/2108.06912
[55] (Pokhrel & Choi, 2020), https://ieeexplore.ieee.org/document/9079513
[56] (Diao *et al.*, 2020), https://arxiv.org/pdf/2010.01264.pdf
[57] https://arxiv.org/pdf/1912.04977.pdf
[58] (Work Programmes - DIGITAL | Shaping Europe's Digital Future, 2021), https://digital-strategy.ec.europa.eu/en/activities/work-programmes-digital
[59] https://digital-strategy.ec.europa.eu/en/news/connecting-europe-facility-2021-2027-have-your-say-cef2-digital
[60] (EUR-Lex - 52020DC0066 - ES - EUR-Lex, 2020), https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593073685620&uri=CELEX:52020DC0066
[61] (Plan de recuperación para Europa, s. f.), https://ec.europa.eu/info/strategy/recovery-plan-europe_es
[62] (Plan de Recuperación, Transformación y Resiliencia Gobierno de España, 2021), https://www.lamoncloa.gob.es/temas/fondos-recuperacion/Documents/30042021-Plan_Recuperacion_%20Transformacion_%20Resiliencia.pdf
[63] (El Gobierno nombra a su primer Chief Data Officer para desplegar la Oficina del Dato en España en el marco del Plan de Recuperación, 2021), https://portal.mineco.gob.es/es-es/comunicacion/Paginas/210720_np_chief.aspx
[64] El Plan de transformación digital de la AGE sin ir más lejos contempla la línea de acción «disponer de sistemas de análisis de datos para la toma de decisiones» en relación con el objetivo «gestión corporativa inteligente del conocimiento, la información y los datos».

REFERENCIAS

AEPD. 2020. «Cifrado y Privacidad III: Cifrado Homomórfico.» AEPD. June 22, 2020. https://www.aepd.es/es/prensa-y-comunicacion/blog/cifrado-privacidad-iii-cifrado-homomorfico.

Al-Rubaie, Mohammad, and J. Morris Chang. 2019. «Privacy-Preserving Machine Learning: Threats and Solutions.» IEEE Security Privacy 17 (2): 49–58. https://doi.org/10.1109/MSEC.2018.2888775.

Chen, Rui, Noman Mohammed, Benjamin Fung, Bipin Desai, and Li Xiong. 2011. «Publishing SetValued Data via Differential Privacy.» PVLDB 4 (August): 1087–98. https://doi.org/10.14778/3402707.3402744.

Cunha, Mariana, Ricardo Mendes, and Joao Vilela. 2021. «A Survey of Privacy-Preserving Mechanisms for Heterogeneous Data Types.» Computer Science Review 41 (August): 100403. https://doi.org/10.1016/j.cosrev.2021.100403.

Danezis, George, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Métayer, Rodica Tirtea, and Stefan Schiffner. 2014. Privacy and Data Protection by Design - from Policy to Engineering. https://doi.org/10.2824/38623.

Izabachène, Malika, Renaud Sirdey, and Martin Zuber. 2019. «Practical Fully Homomorphic Encryption for Fully Masked Neural Networks.» In Cryptology and Network Security, edited by Yi Mu, Robert H. Deng, and Xinyi Huang, 11829:24–36. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-31578-8_2.

Kairouz, Peter, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, *et al.* 2021. «Advances and Open Problems in Federated Learning.» arXiv:1912.04977 [Cs, Stat], March. http://arxiv.org/abs/1912.04977.

Porras Escalada, Diego, Félix Saz Marco, and David Pérez Fernández. 2012. «Desarrollo de Una Aplicación de Código Abierto Para El Control de La Confidencialidad Estadística En La Difusión de Datos de La AEAT.» Revista Estadística Española. vol 54 número 179: 329–55.

David Pérez Fernández. 2014. «Un sistema de cruce de información sensible sin compromiso de la privacidad de los datos fuente». Revista Estadística Española, vol 56, número 185, pp. 279-296