Transferencia de datos cifrados preservando la privacidad en federación de entidades sociales

Septiembre 2021

Con la participación de:

Índice de contenido

Introducción

La publicación de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPD-GDD) así como el posterior Reglamento General de Protección de Datos (RGPD) marcaron un antes y un después en el tratamiento de datos personales.

Los datos personales son el elemento fundamental de trabajo de muchas entidades de ámbito social, recibiendo especial interés aquellas que tienen como foco de su actuación a colectivos vulnerables. Este es el caso de la Coordinadora Síndrome de Down de Catalunya, que nace con el objetivo de trabajar en pro de las personas con discapacidad intelectual, especialmente aquellas que tienen síndrome de Down. La Coordinadora Down Catalunya pretende potenciar, promover, coordinar y atender a las diferentes asociaciones y entidades de iniciativa social que tengan por objeto velar y defender los derechos y la inclusión de las personas con síndrome de Down de Cataluña y de sus familias. Así, la Coordinadora, mediante la realización de actuaciones y actividades, quiere ofrecer un servicio de apoyo a las personas con discapacidad intelectual, a sus familias y a los profesionales.

Down Catalunya reúne a diferentes asociaciones y fundaciones que trabajan en Catalunya con personas con síndrome de Down. Debido a la responsabilidad proactiva de estas asociaciones y fundaciones de salvaguardar los derechos fundamentales de las personas con las que trabajan, sus datos no son compartidos. Esta decisión en pro del derecho a la intimidad de las personas va en detrimento de la capacidad de Down Catalunya de poder describir con detalle cuál es el alcance de las asociaciones y fundaciones que reúne.

Este proyecto nace de la necesidad de Down Catalunya de conocer información básica sobre la actuación de susodichas entidades con el requisito invulnerable de preservar la privacidad de los datos en todo momento.

Actores

En este proyecto participan diferentes organizaciones. Los actores principales son:

- **DataForGoodBCN**, como responsable de las tareas de diseño, gestión, y parte de la ejecución del proyecto;

- **Down Catalunya**, como responsable de definir los requisitos y delinear los objetivos del proyecto; y
- **Universitat Pompeu Fabra**, como responsable del desarrollo de la solución algorítmica así como facilitador de la transferencia de datos.

Además, contamos con la participación de varias asociaciones y fundaciones que reciben el soporte de Down Catalunya. Estas son:

- Andi Sabadell;
- Aura fundación;
- Cromo Suma;
- Down Lleida;
- Down Tarragona;
- FamíliaAMIC; y
- Fundación Astrid-21.

Necesidad

Down Catalunya ayuda a las entidades que reúne a recaudar fondos para incrementar su poder de actuación. Para ello, es necesario que Down Catalunya pueda describir con certeza cuál es el impacto que estas entidades alcanzan así cómo cuáles son sus necesidades.

Para acotar el alcance de la necesidad se acordaron una serie de preguntas que dan respuesta a las cuestiones que Down Catalunya necesita poder abordar. Estas son las siguientes:

Datos genéricos

1. ¿Con cuánta gente con discapacidad intelectual trabajamos?
2. ¿Con cuántas personas sin discapacidad intelectual trabajamos?
3. De las personas con discapacidad intelectual con las que trabajamos, ¿cuántas son socias de alguna entidad?
4. De las personas sin discapacidad intelectual con las que trabajamos, ¿cuántas son socias de alguna entidad?

Datos demográficos y otros

5. ¿Cuántas personas de cada edad participan en alguna entidad socia?
6. ¿Cuál es el número de hombres y mujeres que participan en alguna entidad socia?

7. ¿Cuánta gente participa en cada programa?
8. ¿Cuál es el número total de personas que participan en alguna entidad socia con Síndrome de Down?
9. ¿Cuál es el número total de personas que participan en alguna entidad socia con otra discapacidad intelectual (diferente a SD)?
10. ¿Cuál es el número total de personas que participan en alguna entidad socia por provincia de residencia?

Datos recolectados

Acotar y enumerar las preguntas a resolver nos permite definir de forma adecuada el conjunto mínimo de datos necesarios para obtener la información necesaria. Este ejercicio promueve el principio de minimización desde el diseño.

Los datos que necesitamos recolectar para poder dar respuesta a las preguntas planteadas son los siguientes:

- Nombre y apellidos
- DNI
- Síndrome Down (NO/SÍ)
- Otra discapacidad (NO/SÍ)
- Persona socia
- Fecha de nacimiento
- Sexo
- Programa ofertado
- Dirección postal

Solución criptográfica

La criptografía se basa en usar técnicas de cifrado o codificado para modificar representaciones lingüísticas para hacerlas ininteligibles a personas no autorizadas.

El tipo de propiedades de las que se ocupa la criptografía se pueden resumir en cuatro:

- confidencialidad;
- integridad;
- vinculación; y
- autenticación.

Un sistema basado en criptografía se considera seguro en relación a una tarea si una persona no autorizada no puede romper las barreras de seguridad implementadas.

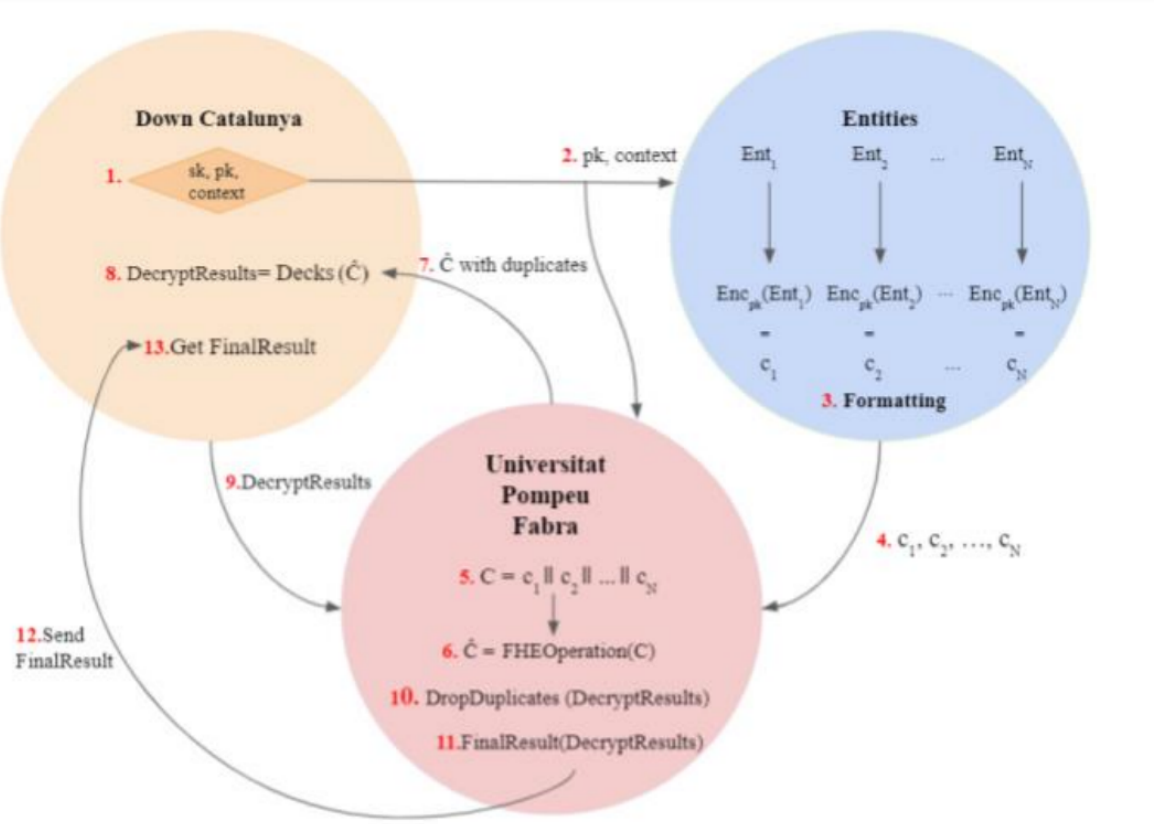
Implementación

El proyecto se desarrolló en Python dada su naturaleza abierta y su uso extendido.

Para desarrollar las técnicas de criptografía homomórfica se usa el paquete PyFHEL, que es el acrónimo para Python for Homomorphic Encryption Libraries. Está construido sobre Afhel, una *Abstracción para Homomorphic Encryption Libraries* en C++, y la versión actual solo soporta SEAL. Además, el proyecto contiene una gran serie de demostraciones, ejemplos, y pruebas para HElib y SEAL.

Arquitectura

El esquema de la solución sigue el flujo propuesto en la siguiente imagen:



El siguiente escrito es una adaptación del trabajo fin de grado de Laia Auset Rizo realizado bajo la supervisión de Sergi Rovira y la Dra. Vanesa Daza en la Universitat Pompeu Fabra[1].

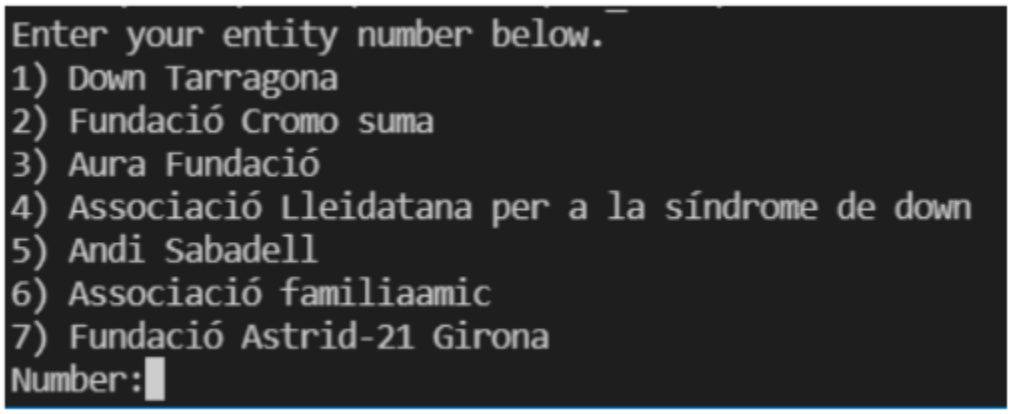
En el paso 1, Down Catalunya se encarga de generar los pares de claves pública y secreta, almacenadas en dos archivos diferentes. También genera una clave de contexto de encriptación homomórfica (EH).

La clave de contexto EH es necesaria para cualquier otra función (cifrado/descifrado y operaciones) y las principales propiedades que se satisfacen es que *p* debe ser un primo y *p-1* debe ser múltiplo de *2*m*, donde *p* es el módulo del texto plano y *m* es el coeficiente polinómico con un valor de *2048*. Si *p* debe ser un primo, entonces debe ser un entero positivo con exactamente dos divisores positivos, 1 y el propio *p* [2].

[1] Auset Rizo, Laia. Fully Homomorphic Encryption for privacy-preserving data analytics. July, 2021.
[2] UPF-MTC 24308 Criptografia I Seguretat (2020, January). Cryptography lecture notes. https://upf-cryptography.github.io/_main.pdf

La clave pública y la clave de contexto se enviarán a todos las asociaciones y fundaciones (a partir de ahora denominadas como AyF) y a la UPF para cifrar y operar con los datos (paso 2), mientras que la clave secreta permanecerá siempre privada por parte de Down Catalunya. En ningún caso se enviará la clave secreta a las AyF, a la UPF, ni a ningún otro tercero. Además, solo se enviará la clave pública y la clave de contexto a las AyF y a la UPF.

En el paso 3, un script de preprocesamiento formatea y encripta todos los datos de las AyF. Al principio, cada AyF tiene un archivo .csv con columnas y filas con todos los datos de cada participante, como nombre, apellidos, fecha de nacimiento, DNI y otros datos personales. Al principio de la ejecución, aparece un menú interactivo básico, y cada AyF escribe su correspondiente número de institución. Este valor nos ayudará más tarde a la hora de identificar a cada persona.



El menú comprueba que la entrada es correcta y pasa a la siguiente funcionalidad. En este punto, el programa comienza a leer el documento de datos de cada AyF y formatea su contenido. Cada documento de cada AyF tiene que tener el mismo formato para que los siguientes programas funcionen. Por ejemplo, el formato de la fecha de nacimiento debe satisfacer la fórmula dd/mm/aaaa, y los campos que contienen nombres y apellidos deben estar en mayúsculas y no contener ningún acento ni carácter especial..

Una vez el programa ya ha validado la fecha de nacimiento pasa a calcular la edad de cada persona.

El aspecto más importante a tener en cuenta es cómo identificamos a las personas. Utilizamos dos identificadores para cada persona, su DNI y otro ID generado por la concatenación de

diferentes valores de la persona usuaria, como el nombre, el apellido, o la fecha de nacimiento al que luego se le aplica una función de hash con el método SHA-512[3]. Para los registros que no tienen DNI, el programa genera un DNI único falso que contiene el número de identificación de la AyF, un contador de ocho dígitos que indexa los registros de las personas de dentro de la AyF, y la letra "I", ya que es una de las letras no válidas en un DNI legítimo. El resultado tiene exactamente nueve dígitos y un carácter, que coincide con la longitud de un documento de identidad válido.

<AyFValue><Counter><Letter”I”>

Por ejemplo, 200000001I, donde el número dos se refiere a la Fundación Cromo Suma, el número 0000001 es el contador del registro de esta persona, y la letra "I". Adicionalmente, se genera el mismo *hashed id* que el primer caso explicado anteriormente con las mismas condiciones.

El contador permite que el falso id identifique de forma única a cada persona dentro de una AyF. Sin embargo, no podemos utilizarlo para comprobar si dos registros de diferentes AyFs corresponden a la misma persona, dado que el falso identificador se construye utilizando un contador y un identificador de AyF que son internos de dicha organización. Por lo tanto, el proceso de formateo también añade un campo a cada registro indicando si el DNI es legítimo o se ha generado en el proceso de formateo. Luego, cuando la UPF elimina los duplicados, utiliza el DNI para comprobar si dos registros corresponden a la misma persona sólo si ambos DNI son legítimos. En caso de que alguno de ellos sea un DNI falso, se usaría el DNI cifrado. En este caso, suponemos que los datos personales utilizados para calcular el hash ID identificarán de forma única a los sujetos de la base de datos. Damos un ejemplo de una posible situación en la que identificamos a una persona utilizando todos los identificadores explicados anteriormente.

Imagina que Alice es miembro de Down Tarragona pero también de Aura Fundació. Down Tarragona es una AyF que sí proporciona el DNI de Alice, pero Aura Fundació no. Cuando cada AyF formatea los datos, en la base de datos de Down Tarragona Alice obtendrá su DNI y hash

[3] Gueron, Shay & Johnson, Simon & Walker, Jesse. (2011). SHA-512/256. 354-358. 10.1109/ITNG.2011.69.

id encriptado. Mientras que en Aura Fundació, Alice obtendrá un DNI falso y el hash id encriptado ya que el campo DNI estará vacío porque la información personal no fue proporcionada por la AyF. En el futuro, todos estos datos se almacenarán juntos concatenando todos los datos de la AyF pero teniendo dos Alice para las dos entidades y contando con Alice como dos personas diferentes. Esto es erróneo.

Gracias a la implementación del identificador explicada anteriormente, este problema puede ser resuelto. En primer lugar, el programa comprueba los DNI de las dos Alice, y observa que un DNI es la identificación válida de Alice y la otra es la identificación falsa de Alice, que contiene dígitos diferentes. En consecuencia, comprueba el hash del DNI de las dos Alice, que acabará siendo el mismo, ya que estos valores contienen el nombre, el apellido, el sexo y la fecha de nacimiento. Aunque los identificadores sean diferentes y el identificador hash sea el mismo, el programa clasifica los dos registros como una sola persona.

A continuación, el programa procede a la tercera validación. En nuestros datos, las personas se clasifican según si han sido diagnosticadas con Síndrome de Down (SD), o con otras discapacidades (OD). La condición de esta última validación debe satisfacer que una persona no pueda estar etiquetada como SD y OD simultáneamente. El siguiente paso es clasificar y eliminar las filas duplicadas de cada documento de la AyF para preparar el archivo a encriptar. A partir de aquí, parte de la propuesta consistió en decidir qué variables debían encriptarse y cuáles no. Las variables que debían encriptarse eran las que contenían información personal o datos sensibles y que un atacante podría utilizar para identificar a una persona:

1. Nombre y apellidos.
2. Documento de identidad (DNI).
3. Fecha de nacimiento.
4. Código postal.

Esta decisión fue considerada junto con diferentes personas expertas en protección de datos. Normalmente, cifrar nombre, apellidos y DNI sería suficiente para proteger la identificación individual, pero siguiendo un criterio de precaución, se sugirió cifrar también y evitar el uso de los campos fecha de nacimiento y código postal ya que, al tratarse de de una población relativamente pequeña, estas variables podrían facilitar la identificación de los sujetos de los datos.

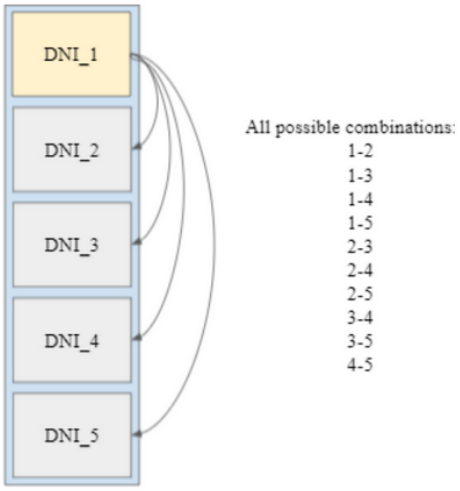
Estos datos se encriptan utilizando la biblioteca Fully Homomorphic, PyFhel. Por último, la salida es un archivo .pickle, ya que nos permite guardar objetos. Estos objetos se almacenan en un diccionario que contiene los datos encriptados y no encriptados de cada AyF.

FormatFile = {DataEncEntity1. pickle, . . . ,DataEncEntity7. Pickle}

En el paso 4, todas las AyF envían sus archivos formateados finales, DataEncEntityX.pickle, a la UPF.

En el paso 5, la UPF comienza a ejecutar un nuevo programa. Primero lee todos los archivos .pickle recibidos de las AyF y concatena todos los documentos formateados y encriptados de cada AyF para procesar un solo documento, llamado dataOperate.pickle.

En el paso 6, el programa comienza a realizar las primeras operaciones con Fully Homomorphic Encryption. Solo necesitamos operar con las variables que contienen información personal y sensible, como el DNI y el *hashed id*. Una de las principales operaciones implementadas es la identificación de duplicados con datos. Lo que hace el programa es comparar cada id con todos los demás, haciendo todas las combinaciones posibles. Resta el primer id (DNI_1) a los siguientes (DNI_2, ..., DNI_5), tal como se muestra en la siguiente figura.



Si el resultado de la resta es cero, significa que los ids son iguales y la persona también. Si el resultado es un valor diferente de cero, significa que los ids son diferentes y ambas son personas diferentes. Sin embargo, estos resultados están encriptados, y no sabemos qué

combinaciones son iguales o son diferentes a cero después de restar, y en consecuencia, no sabemos si hay duplicados.

Una vez realizada la operación, el programa almacena los resultados encriptados de cada sustracción de cada combinación. Además, también necesitamos permutar los datos para garantizar seguridad y privacidad. Si no realizamos este paso, Down Catalunya podría hacer una sistema de ecuaciones y obtener el DNI. El último paso es generar un fichero que se envía a Down Catalunya (paso 7).

En el paso 8, Down Catalunya se encarga de leer este fichero y desencriptar los resultados, ya que es el único socio que tiene la clave secreta. Una vez descifrados los resultados para cada sustracción de id, el programa lee los datos. Luego, comienza a listar las posiciones de los valores que son iguales a cero y los almacena en un nuevo archivo llamado *positionsToRemove.pickle*. A continuación, este archivo se envía de nuevo a la UPF para proceder al proceso de eliminación de duplicados (paso 9).

En el paso 10, la UPF lee el archivo con las posiciones donde los resultados son iguales a cero y puede comenzar a eliminar duplicados. Una vez eliminados, calcula las respuestas a las preguntas propuestas (Necesidad) y genera un fichero de datos final con los respectivos resultados que necesita Down Catalunya (paso 11). Al final, este fichero de datos agregados se envía a Down Catalunya (paso 12) y, en el paso 13, Down Catalunya los recibe.

Impacto

Este proyecto permite la transferencia de datos y la ejecución de un análisis estadístico básico entre diferentes entidades preservando en todo momento la anonimidad de los sujetos de los datos.

Este proyecto es un claro ejemplo de responsabilidad proactiva promovida por el cumplimiento con la LOPD-GDD y el RGPD. Consideramos relevante recordar que el ámbito de aplicación incluye a personas diagnosticadas con Síndrome Down u otros tipos de discapacidad intelectual, así como a sus familias. El colectivo de interés se considera un colectivo vulnerable y los riesgos derivados del tratamiento de susodichos datos son más elevados. La creación de la herramienta presentada responde a este complejo paradigma garantizando los derechos y libertades de los sujetos de los datos.

La ambición de este proyecto es que pueda repercutir en otras entidades con estructura federativa. La experiencia y conocimiento adquirido por el trabajo de DataForGoodBCN nos ha hecho conocedores de otras organizaciones que tienen la misma carencia. La herramienta presentada es lo suficientemente flexible para asegurar su escalabilidad. En consecuencia, esperamos poder transferir esta solución a otras organizaciones para seguir contribuyendo a la cultura de protección de datos así como una utilización responsable de la información de carácter personal conforme a la LOPD-GDD y al RGPD.