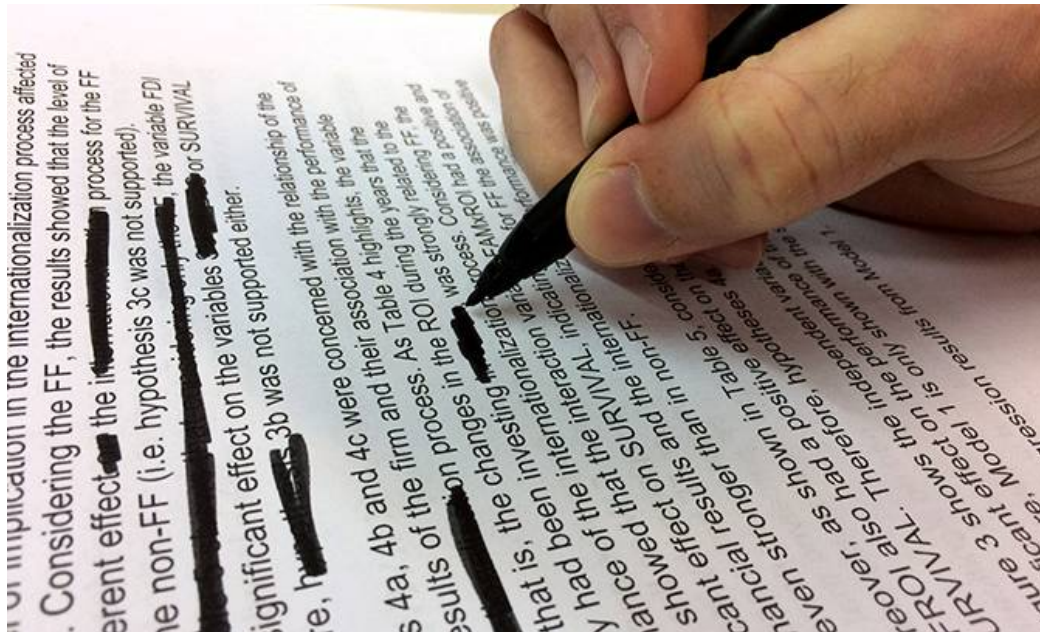


## Cómo ocultar información confidencial de forma automática

Investigadores de la Universidad Rovira i Virgili y la Universidad Abierta de Cataluña han diseñado un sistema que detecta y oculta automáticamente los datos confidenciales que aparecen en los textos, como los de algunos documentos médicos que se pueden encontrar por internet. De este modo, se pueden enviar a terceros sin comprometer la privacidad y manteniendo el anonimato de las personas u organizaciones a las que hacen referencia.

SINC

27/4/2017 12:00 CEST



Hasta ahora, la protección de documentos confidenciales requería de expertos que, manualmente, identificaban y eliminaban palabras. / URV

Actualmente, los datos personales resultan de gran utilidad en muchos ámbitos, tanto de investigación como de negocios o planificación. Por ejemplo, los datos de pacientes almacenados en historiales médicos son fundamentales para realizar investigación médica; las operaciones bancarias constituyen la base de los análisis financieros, y el análisis de las transacciones comerciales sirve para mejorar la personalización de los servicios que se prestan.

Como muchos de estos datos son confidenciales, los documentos que los

contienen deben protegerse antes de ser enviados a los investigadores. Utilizar mecanismos de protección adecuados es, por tanto, esencial para garantizar la privacidad o el anonimato de los individuos.

En este contexto, el investigador David Sánchez del grupo de investigación CRISES-Cátedra UNESCO de Privacidad de Datos, del departamento de Ingeniería Informática y Matemáticas de la Universidad Rovira i Virgili (URV), junto a la investigadora Montserrat Batet del grupo de investigación KISON de la Universitat Oberta Catalunya (UOC), han diseñado un sistema que detecta y oculta de forma automática la información confidencial de documentos de texto.

---

**El sistema analiza documentos disponibles en internet y protege aquellos términos que podrían facilitar deducciones a un tercero**

Aunque la legislación de la Unión Europea es muy estricta en cuanto a la cesión de datos personales sin consentimiento de los implicados, en otros países, como Estados Unidos, es habitual que se requieran y faciliten documentos privados en asuntos judiciales, bajas laborales, seguros, etc. Ahora bien, en todos los casos, se debe garantizar que los documentos facilitados no revelan información confidencial que pueda utilizarse, por ejemplo, con fines discriminatorios.

Hasta ahora, la protección de documentos confidenciales requería de uno o varios expertos que, manualmente, identificaban y eliminaban palabras, sintagmas o frases que podían revelar información sensible o potencialmente discriminatoria. En este proceso se tienen en consideración términos sensibles, como el nombre de una enfermedad contagiosa, y conjuntos de términos que permitirían deducir los anteriores indirectamente, como, por ejemplo, combinaciones de medicamentos o tratamientos que se aplican solo en una enfermedad concreta. Trabajar con estos elementos es una tarea ardua y, por su complejidad, no siempre infalible.

El método que han desarrollado automatiza todo el proceso, lo que permite manejar y proteger de manera eficiente el gran volumen de datos que se

utilizan actualmente en investigación. Para ello, el sistema analiza la información disponible en internet, que es la que un tercero podría utilizar como base de conocimiento para deducir información confidencial de un documento protegido. Seguidamente protege aquellos términos que podrían facilitar tales deducciones.

---

Por ejemplo, en lugar de especificar que un paciente tiene una neumonía, el sistema indicaría que sufre una enfermedad del sistema respiratorio

El nuevo sistema, cuyos detalles se publican en la revista *Engineering Applications of Artificial Intelligence*, permite que los documentos se puedan enviar a terceros sin comprometer la privacidad y manteniendo el anonimato de las entidades (personas, organizaciones, etc.) a las que hacen referencia.

Las pruebas han demostrado que el nuevo método es más exhaustivo y ofrece una precisión equiparable a la de un experto humano. Además, a diferencia de los expertos, el sistema no se limita a eliminar términos peligrosos, sino que, en la medida de lo posible, intenta reemplazarlos por conceptos más generales que resulten más ambiguos.

Por ejemplo, en lugar de especificar que un paciente tiene una neumonía, indicaría que sufre una enfermedad del sistema respiratorio. Con esto se consigue que el documento protegido sea más comprensible y más útil en los análisis posteriores.

Implantado en un prototipo de 'software'

De momento, el método se ha implementado en un prototipo de *software* que se ha probado con documentos clínicos en inglés. Se prevé que en un futuro próximo pueda aplicarse en otros ámbitos de conocimiento y pueda establecerse como una herramienta de uso profesional de especial utilidad para la investigación.

La investigación se enmarca en el proyecto europeo CLARUS, sobre privacidad de datos en la nube, que coordina la URV y que cuenta con la financiación del programa Horizonte 2020 de la Unión Europea para el

período 2015-2017. Asimismo, forma parte del proyecto *SmartGlacis: tecnologías de seguridad y privacidad para ciudades inteligentes* de la UOC, financiado por el Ministerio de Economía y Competitividad.

**Referencia bibliográfica:** David Sánchez y Montserrat Batet: "Toward sensitive document release with privacy guarantees". *Engineering Applications of Artificial Intelligence*. Volume 59, March 2017, Pages 23-34. DOI: [10.1016/j.engappai.2016.12.013](https://doi.org/10.1016/j.engappai.2016.12.013)

Derechos: **Creative Commons**

TAGS

PRIVACIDAD | DATOS | SEGURIDAD |

Creative Commons 4.0

Puedes copiar, difundir y transformar los contenidos de SINC. [Lee las condiciones de nuestra licencia](#)