Recomendaciones sobre datos y software

v 2.0 (febrero 2023)

Los principios de Ciencia Abierta sirven de guía para el uso y re-uso óptimo de datos en la investigación biomédica, ya sean estos datos generados específicamente para este fin o provengan de la actividad asistencial (uso secundario del dato para investigación). Dichos principios también son relevantes para la determinación del conjunto de herramientas necesarias para que la adquisición, manejo, análisis, visualización e interpretación de resultados sea posible. Dada la heterogeneidad de sistemas de manejo de datos, el objetivo primordial es garantizar la interoperabilidad de estos datos, entendida como la posibilidad de usar datos de la misma naturaleza provenientes de distintas fuentes, así como la posibilidad de combinar datos de distinta naturaleza gracias al manejo de ontologías y vocabularios controlados. Este es de hecho un objetivo fundamental de IMPaCT. Los metadatos - información adicional que proveen el contexto en el cual un dato ha sido generado - son tan importantes como los datos en sí, dado que sin esta información de contexto los datos pueden resultar de poca utilidad para desarrollar cualquier investigación posterior. Los metadatos pueden ser de tipo técnico, p. ej. profundidad de secuenciación, o descriptivo, p. ej. comorbilidades del paciente a partir del cual se ha generado dicho dato.

El **programa de Ciencia de Datos de IMPaCT (IMPaCT-Data)** guiado por los principios de ciencia abierta, hace las siguientes **recomendaciones en cuanto a la estandarización de los datos biomédicos para facilitar su interoperabilidad, acceso, uso y gobernanza** a través de sus distintos paquetes de trabajo. Es importante destacar que estas recomendaciones asumen el principio básico de la **federación de datos e infraestructuras** para facilitar la gobernanza de los mismos y una implementación práctica. Estas recomendaciones están en el marco de distintos esfuerzos, proyectos e iniciativas Europeas e internacionales, incluido el trabajo existente alrededor del Espacio Europeo de Datos en Salud (EHDS por sus siglas en inglés), la implementación a través del *European Genomic Data Infrastructure* (GDI por sus siglas en inglés) de las especificaciones acordadas en el proyecto *Beyond 1 Million Genomes* (B1MG por sus siglas en inglés) a partir de la declaración conjunta de los estados miembros y la Comisión Europea de *1+Million European Genomes by 2022* (1+MG por sus siglas en inglés). Estas recomendaciones también se sustentan en el trabajo realizado en proyectos significativos como HealthyCloud y TEHDAS, los cuales se enfocan en distintos aspectos para el uso sistemático de datos en salud para investigación y desarrollo de políticas sanitarias. Finalmente estas recomendaciones se basan en el trabajo de la Alianza Global para la Genómica y la Salud (GA4GH por sus siglas en inglés) a las cuales también contribuye con demostradores asociados a sus estándares.

En cuanto al **software** a utilizar, se recomienda el uso de software con licencias de código abierto que cuenten con su código fuente en repositorios con control de versiones de acceso público, p. ej. GitLab y GitHub. En el caso de que se desarrolle software científico, idealmente se seguirán las 4 recomendaciones de ELIXIR [enlace]. El software deberá estar disponible preferiblemente utilizando contenedores software a través de los canales habituales de la comunidad como Bioconda y Biocontainers, y contar con los metadatos suficientes, p. ej. versión, para asegurar la reproducibilidad de los análisis y el rápido despliegue en sistemas computacionales heterogéneos. El conjunto mínimo de metadatos vendrá dado por los esfuerzos en marcha dentro de ELIXIR, y en concreto, de la ELIXIR Tools platform. En cuanto a los flujos de trabajo, referidos como *workflows*, siguiendo las prácticas actuales dentro de ELIXIR, la recomendación es utilizar el *Common Workflow Language* (CWL por sus siglas en inglés) como lenguaje de especificación y como gestores de *workflows* a: Nextflow, Snakemake y Galaxy. Estos *workflows*

El proyecto IMPaCT-Data (Exp. IMP/00019) ha sido financiado por el Instituto de Salud Carlos III (ISCIII), co-financiado por la Unión Europea, FEDER "Una manera de hacer Europa.

estarán a disposición de la comunidad científica a través de repositorio públicos, siendo WorkflowHub la recomendación actual. IMPaCT-Data facilita en la actualidad las directivas para la progresiva estandarización del software de acuerdo con estas recomendaciones, las cuales serán llevadas a cabo por cada proyecto.

Considerando el **uso secundario para investigación de los datos provenientes de un entorno clínico asistencial**, es necesario considerar la necesidad de establecer una infraestructura física independiente de la infraestructura necesaria para dar servicio a la actividad asistencia. Esta infraestructura independiente, conocidos como lagos de datos, operational data stores (ODS por sus siglas en inglés) o similares, deberán facilitar la ingesta rápida de datos provenientes de distintas fuentes asistenciales y tener suficiente capacidad computacional para realizar el proceso de limpieza, normalización y estructuración (inicial) de la información. Tal estructura de datos, puede contener información clínica, de imagen, p. ej. utilizando DICOM como estándar de información, así como genómica en algunos casos. Dicha infraestructura constituye el punto de partida para el uso secundario en investigación. De hecho, a partir de este punto es necesario llevar a cabo procesos de pseudoanonimización o anonimización, homogenización y control de calidad para garantizar el correcto uso de estos datos en la actividad científica.

En cuanto a la **generación y manejo de datos ómicos de carácter sensible**, estos datos deberán depositarse alternativamente en el nodo central o en la instancia española de la *European Genome-phenome Archive* (EGA por sus siglas en inglés). Como caso intermedio, los datos y sus metadatos asociados pueden organizarse en versiones locales de EGA, que mantendrán el mismo formato y estructura de EGA. IMPaCT-Data ofrecerá las guías y software necesarios para estas instalaciones de estas EGA locales. Para facilitar el descubrimiento y el uso de estos datos ómicos, los datos deben ir acompañados de los datos clínicos asociados, siguiendo los formatos y modelos de datos establecidos por las distintas comunidades. En el caso de las enfermedades raras, el formato recomendado, siguiendo el PoC de 1+MG/B1MG, para expresar la información fenotípica es el uso de términos de la *Human Phenotype Ontology* (HPO por sus siglas en inglés)) y los GA4GH Phenopackets. Para las descripciones asociadas a estudios del cáncer, la información seguirá el modelo de datos previsto en el grupo de trabajo correspondiente de 1+MG/B1MG, que incluye el modelo de datos de la iniciativa ICGC ARGO entre otros.

En cuanto a la **generación y manejo de información estructurada proveniente de observaciones fenoclínicas**, el objetivo es facilitar la interoperabilidad a través del uso de distintos vocabularios controlados y ontologías existentes. En particular: SNOMED-CT, ICD-9 y LOINC, organizados siguiendo modelos de datos ámpliamente utilizados como OHDSI OMOP-CDM. Aunque la organización de los datos clínicos en sí misma está fuera de las competencias de IMPaCT-Data, la conectividad entre recursos se beneficiaría del modelado de la información clínica a partir de la historia clínica electrónica (HCE) basada en OpenEHR, que persiguen facilitar la interoperabilidad semántica y se alinea con los desarrollos existentes a partir de la ISO13606 llevados a cabo por grupos de IMPaCT-Data. De hecho, disponer en los sistemas primarios de modelos basados en arquetipos facilita considerablemente la captura y mapeo de la información desde los sistemas primarios a cualquier sistema destino, ya que la correspondencia se realiza directamente a nivel conceptual *(interoperabilidad semántica)*, y no solo estructural. Cuando esto no sea posible, la disponibilidad de arquetipos normalizados, de la información sanitaria, facilitarán también la captura de información semánticamente ordenada desde cualquier sistema de información sanitario.

En cuanto a **la generación y manejo de información proveniente de imagen médica**, esta seguirá el estándar DICOM que facilita la transmisión y análisis de este tipo de datos independientemente de la

organización interna de cada institución y sus PACS. Idealmente, los datos clínicos asociados a las imágenes médicas deberán seguir los estándares propuestos para manejar la información de carácter clínico mencionados en el apartado anterior.

La **calidad de los datos en los sistemas de origen** es fundamental para que los resultados obtenidos sean correctos, y las decisiones que se tomen a partir de los mismos, útiles. Sin embargo, el contexto en el que se capturan los datos clínicos en el ámbito de la consulta médica es totalmente distinto al de un estudio de laboratorio o un ensayo clínico. Es necesario que los sistemas analíticos ofrezcan un retorno a los clínicos y a los gestores, de manera que el procesamiento secundario de los datos suponga un beneficio directo para los usuarios finales, y éstos comprueben las ventajas de registrar la información de forma adecuada y completa en la Historia Clínica. Este mecanismo de feedback, a modo de "tercera columna" de la Historia Clínica, habilita un diálogo continuo entre los sistemas de uso primario y secundario de datos clínicos, que debe redundar en una mejora de la calidad de la información, y una mejora también de los procesos asistenciales, al disponer de herramientas de ayuda a la decisión integradas en el portal clínico.

En términos de **seguridad de los datos**, IMPaCT-Data enfatiza que los proyectos deberán incorporar desde el diseño el cumplimiento de medidas del Esquema Nacional de Seguridad (ENS; RD 311/2022), con especial hincapié en lo que respecta a la utilización de servicios en la nube y a la identificación de activos, su valoración de acuerdo a las 5 dimensiones que tiene en cuenta el ENS (disponibilidad, confidencialidad, integridad, autenticidad y trazabilidad), el análisis de riesgos y establecimiento de medidas de aplicabilidad especialmente en lo relativo en gestión de accesos. Esto requiere que el componente de Reconocimiento y Autorización (AAI por sus siglas en inglés) sea conforme a las medidas que recomienda el ENS. Todo ello en el marco legal aplicable en materia de protección de datos tal como desarrollan el RGPD y la LOPDGDD.

En cuanto a los **mecanismos de descubrimiento e integración**, entendido este último como el análisis y visualización de datos potencialmente de distintas fuentes, se seguirán los estándares impulsados por GA4GH, ELIXIR y 1+MG/B1MG. En particular, para el descubrimiento de los datos disponibles en las distintas instancias de la red de IMPaCT-Data se propone el uso del GA4GH Beacon en su versión 2.0. En el caso de la información clínica estructurada y de imagen médica, las soluciones a implementar para el descubrimiento de información basado en los metadatos existentes podrán beneficiarse del modelo de datos subyacente del Beacon v2.0 o alternativamente deberán de proveer capacidades similares a las ofrecidas por Beacon. En cuanto a la visualización de los datos y operaciones de integración, estos deben adaptarse a las necesidades de cada comunidad científica, siempre que garanticen la conexión a los recursos de datos y software mencionados en los puntos anteriores. Ejemplos de estos sistemas son los utilizados por distintas comunidades de cáncer como cBioPortal, ICGC-Portal y distintas implementaciones de Molecular Tumour Board (MTBs), una vez adaptados para su uso en el contexto de los proyectos de investigación.

En cuanto al **análisis federado de datos**, que permite reunir cohortes mayores sin necesidad de disponer de repositorios centralizados, es necesario contar con entornos de investigación de confianza (TRE por sus siglas en inglés para *Trusted Research Environment*). Estos entornos suelen basarse en 5 principios de seguridad que tienen que ver con el acceso autorizado del investigador a determinado conjunto de datos en el contexto de un proyecto de investigación para que pueda realizar determinadas operaciones, y que solo los resultados agregados del mismo puedan ser extraídos de dicho entorno. Existen distintas implementaciones de esta aproximación, pero todas ellas deben seguir estos principios.