Recomendaciones sobre datos y software v1.0 (abril 2022)

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.

El **programa de Ciencia de Datos de IMPaCT (IMPaCT-Data)** guiado por los principios de ciencia abierta, teniendo en cuenta las oportunidades que ofrece la interoperabilidad de los datos biomédicos y siguiendo las iniciativas nacionales, europeas e internacionales como la *Global Alliance for Genomics and Health (GA4GH)*, ELIXIR, y la declaración conjunta de los estados miembros alrededor del proyecto *1 European million genomes by 2022 (1+MG*), hace las siguientes **recomendaciones en cuanto a la estandarización de los datos biomédicos para facilitar su interoperabilidad, acceso y gobernanza** a través de sus distintos paquetes de trabajo.

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.

En cuanto al **software** a utilizar se recomienda el uso de software con licencias de código abierto que cuenten con su código en repositorios de control de versiones, 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 preferiblemente contenerizado, y disponible 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 set 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, *pipelines*, la recomendación es utilizar lenguajes de especificación de *workflows* como CWL y gestores de *workflows* con los siguientes sistemas recomendados: Nextflow, Snakemake o Galaxy. Estos pipelines estarán a disposición de la comunidad científica a través de repositorios de workflows con la recomendación actual de *WorkflowHub*. IMPaCT-Data facilitará los instrumentos y directivas para la progresiva estandarización del software de acuerdo con estas recomendaciones que llevará a cabo cada proyecto.

En cuanto a la **generación y manejo de datos ómicos de carácter sensible**, estos datos deberán de depositarse en el nodo central o la instancia española de la *European Genome-phenome Archive (EGA)*. 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 de terminos de HPO (*Human Phenotype Ontology)* y GA4GH phenopackets. Para las descripciones asociadas a estudios del cáncer, la información seguirá el modelo de datos previsto en ICGC-Argo o en el grupo de trabajo correspondiente de 1+MG/B1MG.

En cuanto a la **generación y manejo de información estructurada proveniente de observaciones fisioclinicas**, 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 ampliamente 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.

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.

En términos de **seguridad de los datos,** IMPaCT-Data enfatiza que los proyectos deberán incorporar desde el diseño al cumplimiento de medidas del Esquema Nacional de Seguridad (RD 3/2010), 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. 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 deberán de proveer capacidades similares a las ofrecidas por Beacon. Las implementaciones concretas están siendo discutidas en grupos de trabajo y proyectos internacionales en estos momentos, en particular en el contexto de los grupos de trabajo de EHDS/TEHDAS, 1+MG/B1MG y HealthyCloud, de cara a su futura implementación en el EHDS. 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 IMPaCT a los estándares de datos y software mencionados en este documento.