Requirements for Meta(data) Catalogue Software Selection

Table 1

List of requirements

R1	Open-source solution
R2	Preferred are tools implemented in genome-related or medical-related use cases
R3	Eligibility for sensitive data handling
R4	The possibility to authenticate potential requestors via the OAuth protocol
R5	Data catalogues under the Apache License V2.0, Massachusetts Institute of Technology (MIT), or Berkeley Software Distribution (BSD) license are favoured over GNU General Public License (GPL) family licenses
R6	Scalability
R7	Preferred are tools requiring the least effort in terms of technical development
R8	Highly customizable solution ideally supporting FAIR Principles
R9	Software built on up-to-date technology
R10	Supporting PostgreSQL relational database
R11	Supporting ElasticSearch or SOLR search service
R12	FAIR-friendly tool
R13	User-friendliness
R14	The possibility to filter several parameters at once
R15	Software compatibility with FAIR Data Point

Comments on requirements

Before the review was performed, system selection requirements were defined. To satisfy the demands, a suitable catalogue should be open source (R1) with the possibility of customization. Preferred are tools that were implemented in genome-related or medical-related use cases (R2). The ideal solution must be eligible for sensitive data handling (R3), and it should provide the possibility to authenticate potential requestors via the OAuth protocol (R4). Due to the nature of data, this is especially important because MMCI management wants to be sure it makes vulnerable data available for proper purposes.

The next requirement is connected to the license under which the tool is available. Data catalogues under the Apache License V2.0, Massachusetts Institute of Technology (MIT), or Berkeley Software Distribution (BSD) license are favoured over GNU General Public License (GPL) family licenses (R5) due to their permissive character (FOSSA Editorial Team, 2021). GNU GPL (Free Software Foundation, 2022a) licenses are stricter in preventing proprietary commercialization because it requires that any derivative work must be distributed under the same or equivalent license terms. It includes an obligation to release complete source code and all rights to modify and distribute the entire code. On the contrary, Apache License V2.0, MIT, or BSD family licenses impose minimal restrictions on the software's future behaviour. Those licenses allow releasement of the code under any license, including proprietary usage. Somewhere between stands Lesser General Public License (LGPLv3) (Free Software Foundation, 2022b), which is a part of

GNU GPL family licenses, but it provides a more permissive alternative. Unlike GPL, it enables the user to mix the open software with the non-free one. Although the aim is not to create proprietary software, permissive licenses are considered a better option since they do not impose any restrictions on software usage. Moreover, they do not obligate the user to publish every single change performed on the software, which can be limiting to MMCI technicians.

Since the solution needs to be suitable for the long term, it must be scalable (R6). Scalability is an essential feature of the proposed solution because it ensures that the data storage method will be applicable for a long time without adding extra space or running out of unique identifiers. In order to better understand the increase in the amount of data per year, the number of sequenced patients in 2021 was examined. Laboratory staff informed that NextSeq 550 machine had completed 21 runs, each with 8 patients sequenced. MiSeq had completed two different types of runs with varying kits of sequencing - 12 runs, each run with 24 patients sequenced, and 19 runs, each with 12 patients sequenced. This information is clearly displayed in Table 2. The total number of sequenced samples at MMCI machines in 2021 was 684. Thanks to this analysis, it is obvious that it is enough to consider the increase of up to 1000 samples per year for further data storage development.

Table 2

Number of Sequenced Samples Per Year 2021

Sequencing machine	Library preparation kit	Number of runs per year 2021	Number of sequenced samples per run	Total number of sequenced samples per year 2021
NextSeq 550	TruSight Oncology 500 assay	21	8	168
MiSeq	Accel™ Amplicon Custom Core Kit	12	24	288
MiSeq	KAPA HyperPlus Kit	19	12	228

Note. Information displayed in Table 2 was obtained from the laboratory staff of MMCI.

For this work, the data catalogue which meets defined requirements the best is chosen from the available open-source software. The aim is to find the solution that requires the least effort in terms of technical development (R7). That means it is undesirable to design own software from scratch, on the contrary, it is preferred to use a freely available solution that best fits sequencing (meta)data publication while it is in accordance with FAIR Principles and highly customizable (R8).

The ideal software should be built on up-to-date technology (R9). It means preferred are technologies with active development over obsolete solutions such as Pascal or Perl. Ideally, chosen tool should support relational database PostgreSQL (R10) and Elastic Search or SOLR search service (R11).

Since the overall goal of this thesis is to FAIRify sequencing data, the related requirement is to build a solution on FAIR-friendly tools (R12).

Among the last but also very important requirements belong user-friendliness (R13). In a selected system, data should be imported easily, and users can efficiently filter many parameters (R14). As stated in the section devoted to the FAIRification process (see subchapter 2.4), software compatibility with FAIR Data Point (R15) represents an advantage, however, it is not a necessary requirement. The FDP is a new concept, which is currently integrated only into several tools acting like data points (FAIR Data Point, 2022).