945358-BIGPICTURE
Central Repository for Digital Pathology

WP2 - Infrastructure and database hosting

D2.05 - Report on pilot operations

Lead contributor	2 - LIU
	joel.hedlund@liu.se
Other contributors	5 - CSC 6 - UU 37 - JANSSEN 38 - BAYER

Document History

Version	Date	Description
V0.1	19 Dec 2023	First Draft
V0.2	22 Dec 2023	Comments
V1.0	13 Jan 2023	Second Draft
V1.1	24 Jan 2023	Final Version

Abstract

This report describes experiences from task T2.5 Pilot operation in WP2 Phase III Refinements, from operating the Bigpicture repository pilot system for one year, implementing successive adaptations to needs in digital pathology as identified by other Bigpicture work packages.

The WP2 activities are carried out primarily by distributed scrum teams at CSC, LiU and UU, running a central repository infrastructure composed of two data centers that are operated by CSC and UU, in an approach that offers high availability and disaster resilience, as well as increased range of services accessible to repository users.

Task T2.5 Pilot operation focused on targeted data incorporation in the repository, in collaboration with pilot data contributors in WP3, through user interfaces and APIs developed in task T2.3 Pilot system development.

At the early stages of pilot operations, the infrastructure was set up to cater primarily to advanced users with a high degree of technical competency, with the intention to build and improve upon repository interface tooling to make it more accessible for a broader range of users and competencies, in task T2.6 Iterative development. The APIs provided by this tooling are available for use by frontend developers in WP4, should the need for further specialized frontend user interfaces arise.

For task T2.5 Pilot operation, storage capacity was procured to receive 0.5 PB slide data, which proved sufficient for this task. Compute capacity was also installed to carry out inference against stored data at pilot scale, which was done in collaboration with pilot inference model contributors in WP4 using models compatible with the format specification described in D4.05.

The time frame of this task was set so as to minimize the risk that unforeseen events in preceding tasks would cause delays in subsequent tasks, however no such events came to pass.

Data submission and transfer services as well as storage backend are operational.

Methods

The Bigpicture repository infrastructure is operated by the two data centers by CSC and UU for high availability and disaster resilience.

WP2 organizes weekly meetings at different operational levels, where interested parties from all work packages are invited to join. The topical ad-hoc cross work package meetings had addressed specific challenges, such as adapting and deploying software platforms such as Cytomine and grand-challenge.org in the Bigpicture repository infrastructure. WP2 is organising an architecture taskforce, which has been focusing on the various processes during the second project year and WP2 representatives are also participating in the Bigpicture honest data broker-, metadata-, and GDPR task forces.

During the course of task T2.5, WP2 arranged its second annual workshop, as a 24 hour, lunch-to-lunch hybrid meeting at the LiU Center for Medical Image science and Visualization (CMIV), which was attended by participants from 14 partners, 8 countries, and 5 work packages (WP1, WP2, WP3, WP4, WP5). WP2 also arranged an internal hackathon workshop for repository infrastructure developers, as an (almost) fully physical lunch-to-lunch meeting at CSC. These have been organized as hybrid meetings. As secure processing of sensitive personal health data for research has been expanding with new projects, the WP2 partners have carried out multiple recruitments and reorganizations in this period to meet the expanding needs.

Results

The services

The two data centers that together form the Bigpicture central repository are operated by CSC and UU. Data submission and transfer services as well as storage backend are operational. CSC and UU offer a bit different services for use in Bigpicture. For example, UU is offering up Cytomine server as-a-service, for use in Bigpicture. Similarly, CSC is offering up services from its expanding sensitive data services suite for use in Bigpicture, such as the secure remote desktop solution SD Desktop, which has been audited against the Finnish national standard for secondary usage of health data.

During T2.5 Pilot operation the basic data submission interfaces have been operated and there has been various development activities (T2.6 Iterative development) continuing related to data mirroring and metadata submission, discovery service as well for the data access service offering. Data mirroring between CSC and UU was originally implemented as a semiautomated process, in order to better learn what the operational parameters should be, before solidifying them into a fully automated process. The intent is to have the first version of the automated process in operation in 2023 Q1. Meanwhile data can be mirrored manually.

Pilot datasets

For task T2.5 Pilot operation, storage capacity was procured to receive 0.5 PB slide data, which proved sufficient for this task. Two data collection milestones were scheduled to this period, and met. The first one, MS5 - Significant amount data received (approx. 75TB) was scheduled to M15, as a first verification of technical hardware scalability. This was also designed as a stepping stone in preparation for milestone MS8 - Delivered clinical dataset via the data submission portal to the BIGPICTURE repository, which was scheduled to M18.

For MS5, 76.5TB of pilot data was received by the repository, including clinical and non-clinical WSI and metadata from Bigpicture beneficiaries from academia, caregivers and EFPIA companies. The pilot data included Bigpicture standards compliant image and metadata formats, as well as non-compliant data and mock-up data, all thoroughly anonymized or even randomly generated in order to avoid risk of privacy protection issues.

For MS8, an addition of three datasets containing clinical WSIs and associated Bigpicture standards compliant metadata were submitted to the repository by Bigpicture partners Region Östergötland, Medical University of Vienna, and University Medical Center Utrecht. With those datasets 78 TB data has now been stored (as per public metrics).

Data Submission interfaces

At the early stages of pilot operations, the infrastructure was set up with to cater primarily to advanced users with a high degree of technical competency, with the intention to build and improve upon this repository interface tooling to make it more accessible for a broader range of users and competencies, for example extending it with user interfaces developed in WP4, during task T2.6 Iterative development. However, such extensions proved not to be needed in the pilot operation phase, as the original tooling set up by WP2 was seen as fully adequate for purpose by the pilot users.

The data submission interface at UU is implemented as a customer-separated object storage endpoint, and a command line user interface sda-cli[1] is provided to streamline the process for the particularities

[1] https://github.com/NBISweden/sda-cli

and rigors of handling and transmitting potentially confidential encrypted data over the public Internet (see also D2.04 Report on status of pilot system implementation). This interface was used by Region Östergötland, Medical University of Vienna, and University Medical Center Utrecht to submit clinical datasets to meet milestone MS8. The feedback from the data contributors regarding the submission experience was exclusively positive, although this command line interface was not explicitly made to be attractive to clinical personnel, but was rather intended to be a low-level user interface primarily targeted to data scientists and system administrators. Nonetheless, WP2 still works to extend the submission tooling for improved accessibility, working under the assumption that the pilot data contributors may represent a subset of the envisioned data contributor community that has an unusually high degree of in-house available technical expertise, and that further accessibility features may be needed to appeal to the full range of data contributors which Bigpicture hopes to engage.

CSC offers an SFTP based data submission interface, however is planning to migrate this protocol over to use SD Connect, which (similar to the UU solution) is using customer specific object storage as an inbox directly. In the current design, this capacity is tied to each customer, who then must get a CSC user account and project.

CSC has been developing the SD Submit service for metadata submission. It provides an API, where also Deciphex data submission tool will integrate. In the future there can be a user interface, but the main method is to use XML files, which are exported from the data submitter's internal systems.

The APIs provided by the submission interfaces developed by WP2 are available for use by frontend developers in WP4, should the need for further specialized frontend user interfaces arise.

Computing services

Compute capacity was also installed to carry out inference against stored data at pilot scale, which was done in collaboration with pilot inference model contributors in WP4 using models compatible with the format specification described in D4.05. The pilot inference uses the initial version of the private inference execution model developed in task T2.3, intended for further refinement in task T2.6.

Grand-challenge.org, a machine learning platform developed at Radboudumc, software platforms' integration work to data repository continues. A pilot Cytomine instance is running in Sweden. UU has been piloting use of Cytomine and grand-challenge.org software on virtual machines in the same system, which would be used for production.

WP4 will develop it further e.g. authentication and authorisation framework. The grand-challenge.org software platform is being investigated for internal use, but is not yet offered up for pilot use. This is not seen as an urgent priority, as it is not immediately intended for use in data submission nor as a primary tool for direct access to AI models or datasets, but is pursued as a means to facilitate indirect access to AI models and datasets.

Security

In parallel to task T2.5, and of tangential relevance to this report, WP2 has started activities in Task T2.7 Regular security review. Here, annual security review will be carried out with regard to components where changes have been made, and penetration testing of the integrated system will be carried out in collaboration with independent external cyber security expertise. This work has been set up with a similar structure as in Task 2.4 Security review of pilot systems. The services of an external penetration testing company was procured by CSC during 2022 Q4, and the penetration testing exercise is scheduled to conclude before M24.

However, in regards to privacy safeguards, and of relevance to this report, WP2 has contributed to

authoring the Technical and Organizational Means deliverable document (D5.4), led by WP5. The intention is that data contributors will be able to use this document to evaluate whether the protective measures that are put in place are sufficient to safeguard the privacy of their data subjects should their data be submitted to the Bigpicture services.

Conclusion

During T2.5 Pilot operation the basic data submission interfaces have been operated and there has been various development activities (T2.6 Iterative development) continuing related to data mirroring and metadata submission, discovery service as well for the data access service offering. UU has been piloting use of Cytomine and grand-challenge.org software on virtual machines in the same system, which would be used for production. There is integration work still needed. Security penetration testing of the integrated system will be carried out at the end of second project year in collaboration with independent external cyber security expertise.

During the pilot phase was received ca. 76 TB anonymous or non-clinical data. It was submitted with command line tools. Metadata model will improve and thus resubmission of the datasets is expected and naturally more and more datasets. MS9 i.e. submission of non-clinical dataset preparation with WP3 and WP4 is on-going. This includes integration with the data submission APIs.

Repository infrastructure pilot operations have been largely uneventful, and WP2 is ready to meet the anticipated accelerated pace of data collection as the repository enters Phase IV Large scale archive operations from M25 onwards.