Workshop Feedback

General Feedback

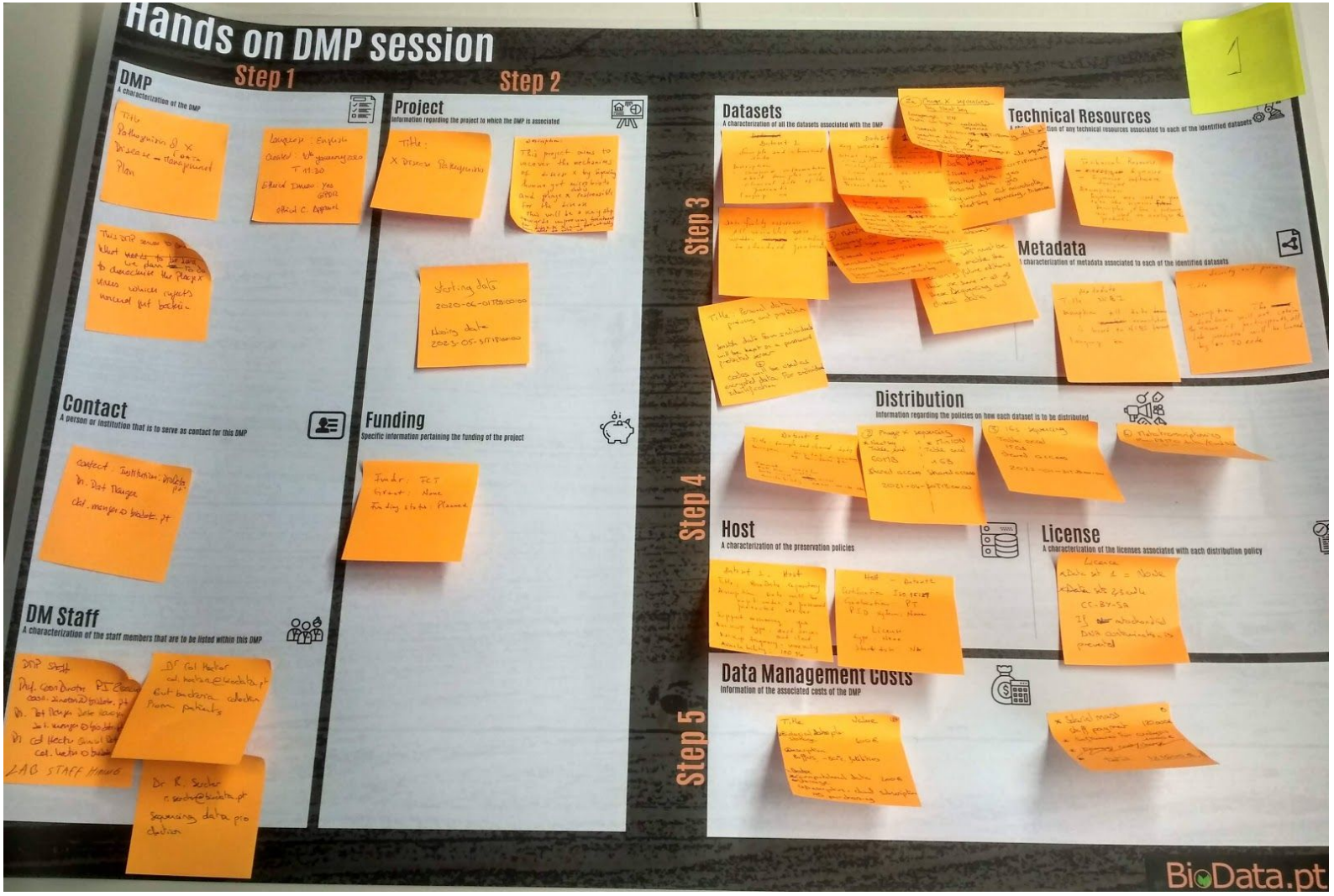
Overall, all groups showed engagement with the Hands-On Data Management Plans exercise, devoting themselves to the task of understanding the data management issues underlying the research project and detailing them in their data management plan. There were some interesting discussion points raised throughout the exercise, including the topic of data privacy and the risk of identifying patients through gut metagenomics due to contamination with mitochondrial DNA.

In general, the information that all groups listed in their data management plan was correct and fairly adequate. Sections 1 and 2 were nearly indistinguishable between the groups, as they all closely followed the project. The main differences were in the level of depth and detail provided in sections 3 and 4.

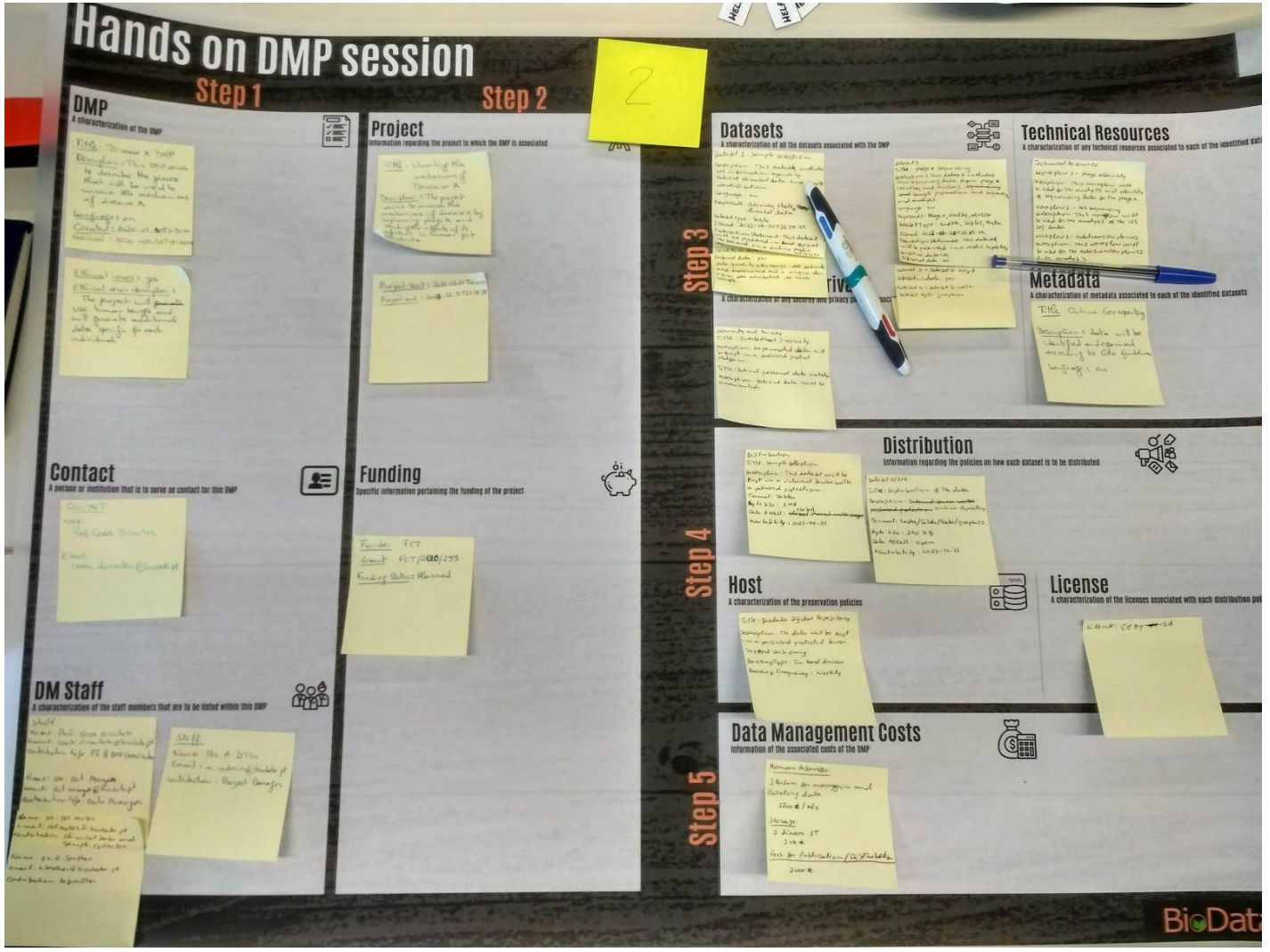
A few key issues were common to several groups

- The use of proprietary encrypted file formats (e.g. .xls) for dataset publication—open source plain file formats are the best practice for findability, interpretability, and long term accessibility.
- Preference for non-commercial usage licenses for their datasets. This is unrealistic in a project related to healthcare, as it might preclude pharmaceutical companies from applying the project's results into solutions for the disease. It also violates the spirit of public funding for scientific research, which is given in the expectation that scientific discoveries will give back to the economy. Unless researchers plan on leveraging their discoveries themselves to start a company (in which case they may want to delay or abstain from publishing their data) they should not restrict others from doing so.
- Identification of only computational resources in the technical resources section, leaving out other critical resources such as sequencing machines.
- Identification of only storage costs in the costs section, leaving out other key costs that apply in this context, such as personnel and material.

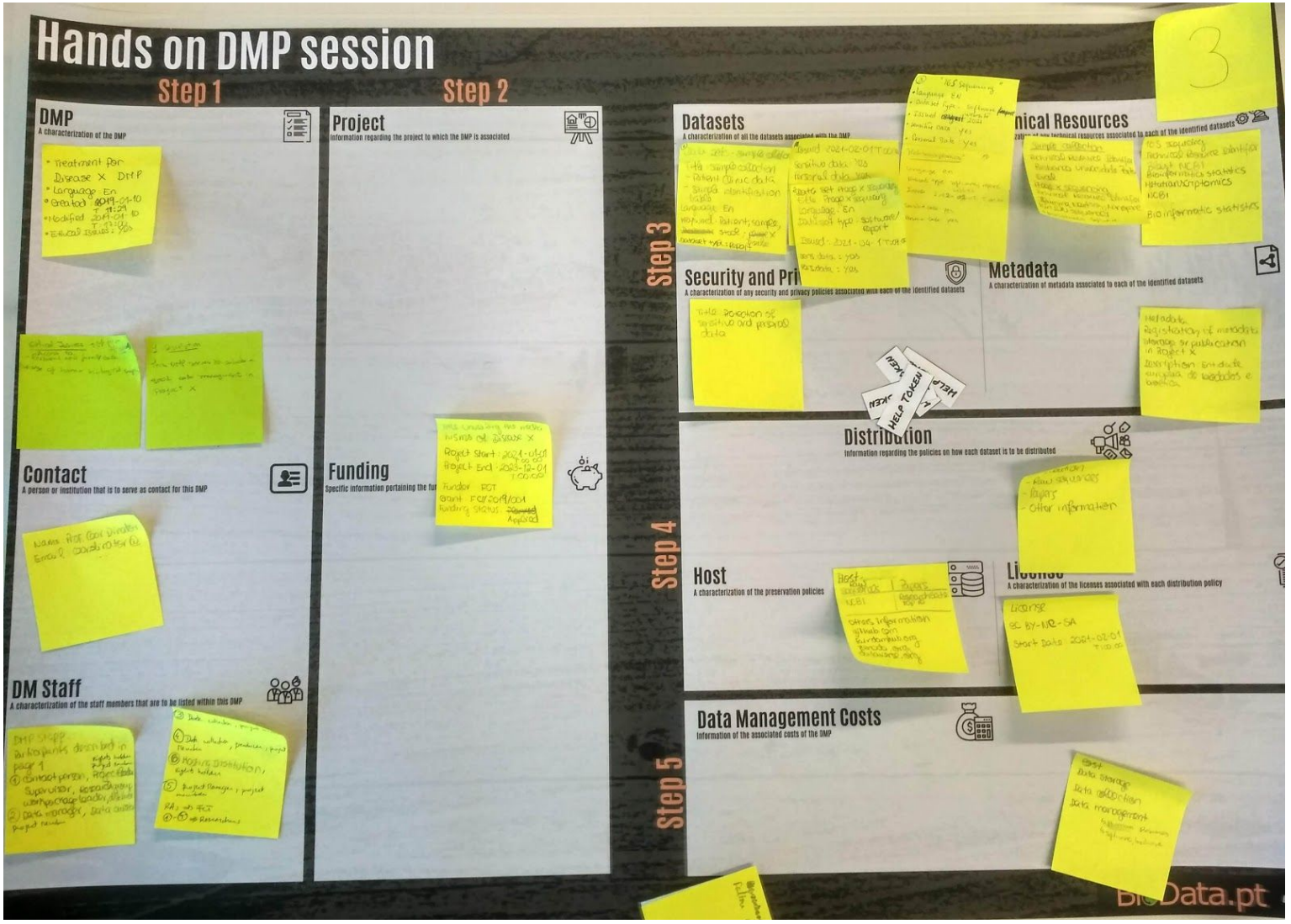
Group 1's DMP was fairly correct overall and had a very section 3. Thee two main issues were: listing only sequencers in the technical resource identification, which should include all software and hardware; and the use of proprietary data formats for dataset storage, which limits long term preservation.



Group 2 created a rich DMP with depth and detail in all sections. Their strongest points were identifying publication fees in section 5 and choosing an open licence. However, their metadata and distribution policies were too simplistic, and they failed to identify the technical resources correctly.



Group 3's DMP was fairly accurate, but the least rich overall, as they added little depth to the information listed in the project. Sections 3 and 5 were the strongest points in this DMP, with the weaker points being the selection of a non commercial licence and the basic distribution policies.



Group 4 created the closest representation of a real DMP, with a consistently high level of depth and detail in all steps. Their strongest points were their rich dataset characterization (including data cleaning) and distribution policies, and their only major issue was the choice of a non-commercial licence.