TOOLBOX DIALOGUE INITIATIVE ™

Toolbox Workshop Report
NMDC Workshop: Linking MIxS standards, Environment Ontology, & GAZ
October 21, 2019
Report submitted January 21, 2020

Prepared by Stephanie E. Vasko, Marisa A. Rinkus, and Chet McLeskey

1. Workshop Summary

The Toolbox Dialogue Initiative (TDI) conducted a workshop at the NMDC Workshop: Linking MIxS standards, Environment Ontology, & GAZ on October 21, 2019 in Berkeley, CA. This workshop involved three dialogue sessions in which 31 people participated. The workshop took place over three hours with dialogue structured by a 25-prompt instrument. The instrument was designed for participants by the TDI team, in consultation with the representatives from NMDC Aim 4, including the Aim 4 lead.

The workshop began with a brief preamble covering the Toolbox approach, instrument design, and details about the workshop. After the preamble, the workshop participants were directed to the TDI app to complete the pre-dialogue survey-like instrument. They then engaged in dialogue sessions that lasted approximately 75 minutes and completed a post-dialogue instrument. After this step, participants were guided through a co-creation activity focused on what makes “good” and “bad” data sets and what an ideal data set might look like.

2. Dialogue Sessions

The main themes from the dialogues were open science, metadata, incentives, timescales for this type of work, and possible deliverables, although the discussion groups varied in their coverage of these themes. The dialogues were not balanced evenly among all participants, but the facilitators attempted to give time and space for everyone to contribute.

The first dialogue group began with a discussion of open science. While they initially thought that they agreed on what open science was, they realized upon subsequent discussion that open data and open science are more nuanced, may depend on the audience, and that open science and open data may be philosophies rather than concepts with singular definitions. A large part of their discussion was around metadata curation, including the ownership/credit of the metadata collection, the quality of the metadata collection, the interoperability of metadata, and the realities versus practicalities of prework versus field work. The group wondered about measuring metrics on data. They wondered if metadata, citation, and downloads are linked, how to determine who gets credit for metadata, and how that translates into retention, promotion, and tenure. The group agreed on the specific challenges, but noted that there is a difference between this community and external communities. They underscored that it is important to recognize different understandings by policy representatives, funders, and journals, but that this is a lot to ask the broader community to keep track of. This group used the prompts to focus on activities or ideas that the NMDC could operationalize. This team also speculated on future activities and roles for the NMDC. Some of the suggested roles for the NMDC would be to educate funders and journals, to think about how to incentivize data and metadata collection, and to develop guidelines and checklists for journals and reviewers to use in evaluating articles. This team also discussed how to have equal voices in conversations.

The second discussion group bounced from module to module, spending the majority of the time on Open Science & Standards, Communication, and Collaboration. Prompt 7 in the Collaboration Module “I am incentivized to collaborate

within NMDC" was where the discussion began as participants felt that collaboration and data management were not incentivized. In particular, granting agencies should ask for it or researchers should be penalized for not doing it. In this group, there was general agreement regarding the necessity and importance of open science. However, participants noted resistance to data sharing as "people are afraid" of getting scooped on an article and that there is a general belief that "this is their data...but they are the caretaker of the data." This quote referred to the fact that people think the data belongs to them, but really, they need to think about themselves as "caretakers" of the data. There was also the realization that early career researchers need to be protected by allowing them to have exclusive rights to their data for six to nine months before being required to make it available to other researchers. This led to more specific discussion of data management and the time that has been invested in collecting and curating data, and the incongruence between older repositories and newer repositories. Through the dialogue, participants identified several important factors for building a database collaborative that would garner buy-in for collaboration. These included: transparency, user-centered focus, easy submission process, ensured data quality, availability of resources, and longevity.

The third discussion discovered that many of the issues they addressed were connected in ways they had not noticed before the dialogue. Beginning with what it means to be "open" in science, they agreed that advancing knowledge and making it available to others is a good starting point. Lowering the bar for data access and other forms of scientific understanding ought to be a goal, but simply having access to a computer is not enough. The group discussed concerns regarding paywalls, publication practices and incentives, and the difficulty of explaining scientific findings to the public. All of these present challenges that the NMDC ought to address systematically. Related to this, they discussed the need to allocate resources to making data interoperable and readily available. Group 3 found collaboration to be a tough balancing act involving expertise and inclusivity. Being at different career stages and from different disciplines can complicate matters when addressing problems in all phases of research. Another potentially complicating factor involves the funding of projects; for example, a given funding source may privilege some forms of expertise over others (either implicitly or explicitly). The solution, they found, seems to lie in the environment of the team—it is crucial to empower all team members to contribute at all phases of the project. In the end, the group acknowledged that there are many social and cultural aspects to the problems that teams face, involving values, incentives, and power structures, and that any effort in open science must address these as directly as possible.

3. Pre/Post Instrument Scores

Figures 1-4 in Appendix 2 provide a visual representation of the pre and post responses for each prompt by module using box and whisker plots. The "box" represents 50% of all responses and the upper and lower "whiskers" each account for 25% of all responses. The longer the box, the more variability in the responses across participants, and vice versa. In some cases, the absence of a response or a "don't know" response accounts for the longer "whisker."

Overall, the Communication module exhibited majority agreement among participants (prompts 1, 3, 4, and 6); however, it should be noted that in some cases there were outliers. There also appears to be general agreement on the following prompts: Open Science & Standards 1, 2, and 3, Data 6, and Collaboration 2, 3, and 6. The most variability was evident for prompts from the Data module, in particular prompts 1, 2, 3, 4, and 5.

4. Co-Creation Activity

The co-creation activity provided participants the opportunity to reflect on the nature of data sets in a way that was informed by the dialogue, and also primed the group for the rest of the NMDC workshop activities. They were first asked to spend five minutes individually brainstorming what makes the data set they submitted (or a data set in general) "good" or "bad." They were then invited to share these results in small groups and create lists of these qualities on large post-its. Each discussion group then came back together and used these "good" and "bad" lists to create a post-it describing the characteristics of an ideal data set. Each discussion group then shared their final activity with the full workshop group.

The results of this final activity are included in Appendix 3 and have informed our recommendations.

5. Recommendations

Based on data gathered from the workshop participants, we offer the following recommendations for your consideration:

- *Create a definition or set of definitions for "open science."* All three discussion groups covered this topic and noted various definitions for the term. Given this, a collective set of definitions of the terms 'open data' and 'open science' might help NMDC frame internal and external discussions. It might also be useful to create a list of the different definitions for these terms based on field and venue (e.g., journals, funding agencies).

- *Develop ways to measure metrics on data sets.* This might help determine if the inclusion of more metadata leads to more citations and/or more downloads. Metrics may also include how to determine who gets credit for metadata and how that translates into retention, promotion, and tenure.

- *Work with communities to incentivize open science and the collection of metadata*. This could be done by working with journals on credit and recognition (like covers), developing awards, and creating additional CV-able recognition for curators. This folds into the metrics recommendation above vis-à-vis retention, promotion, and tenure.

- *Underscore the temporality of the work.* Many of the groups discussed that this work will be evolving and that this stage could just be the beginning stage. Thinking about how the NMDC and the NMDC's activities could evolve might be helpful for the communities involved.

- *Continue having large group discussions and meetings such as this one.* These discussions help build capacity and community among participants and allow members to share perspectives, experiences, and ideas.

6. Conclusion

This workshop sought to generate the following outcomes: (1) enable participants to talk to each other about their work and their work with the NMDC; (2) give participants the opportunity to get to know one another in a deep and coordinated way around the issues that matter for this project; and (3) prime the participants for the activities in the rest of the NMDC workshop.

We believe that (1) and (2) were achieved, and based on the enthusiasm for the co-creation activity, we believe that participants were also primed for subsequent activities and discussion, however, we would rely on the NMDC team for confirmation based on their impressions of the rest of the workshop.

Appendix 1 – Toolbox Dialogue Initiative Instrument

Disagree *Agree*
1 2 3 4 5 I don't know N/A

Open Science & Standards

Core Question: *What are the specific challenges of contributing to open science and standards?*

1. Open science is necessary for advancing knowledge.
2. Consistent data collection and curation is required for effective open science.
3. Metadata is an unnecessary burden on researchers.
4. Sample collection methods can never meet the same set of FAIR data standards.
5. Making all microbiome data interoperable is impossible.
6. We all agree on the specific challenges of contributing to open science.

Data

Core Question: *How do we collect and analyze data?*

1. Bias is unavoidable in data collection and use.
2. Following all the data protocols all the time is not possible.
3. I am aware of the barriers to adoption and compliance for FAIR data.
4. We agree on what it means to annotate data well.
5. We agree on what it means to make genomic data discoverable.
6. Outreach is just as important as data collection and analysis.

Communication

Core Question: *How important is communication to successful collaboration?*

1. Successful collaboration within NMDC requires clear, detailed, and regular communication between members of the team.
2. Your published research isn't finished until it is communicated to the public.
3. Your published research isn't finished until your curated data are made available.
4. Community engagement that supports open science and shared ownership is key to NMDC.
5. Our different domain area experts can communicate across the science, ontology, technical, and community domains.
6. Data management is an aspect of research communication.

Collaboration

Core Question: *What are the specific collaboration challenges we might encounter?*

1. It is clear how we will measure the success of our collaboration.
2. Members of the team must share a common understanding of their project's key concepts.
3. Each discipline involved in this project is equally important to the project's success.
4. Everyone on this project has an equal voice regardless of their career stage.
5. Everyone on this project has an equal voice regardless of their disciplinary background.
6. Interdisciplinary research is more likely to foster innovation than disciplinary research.
7. I am incentivized to collaborate within NMDC.

Appendix 2 – Pre/Post Instrument Data

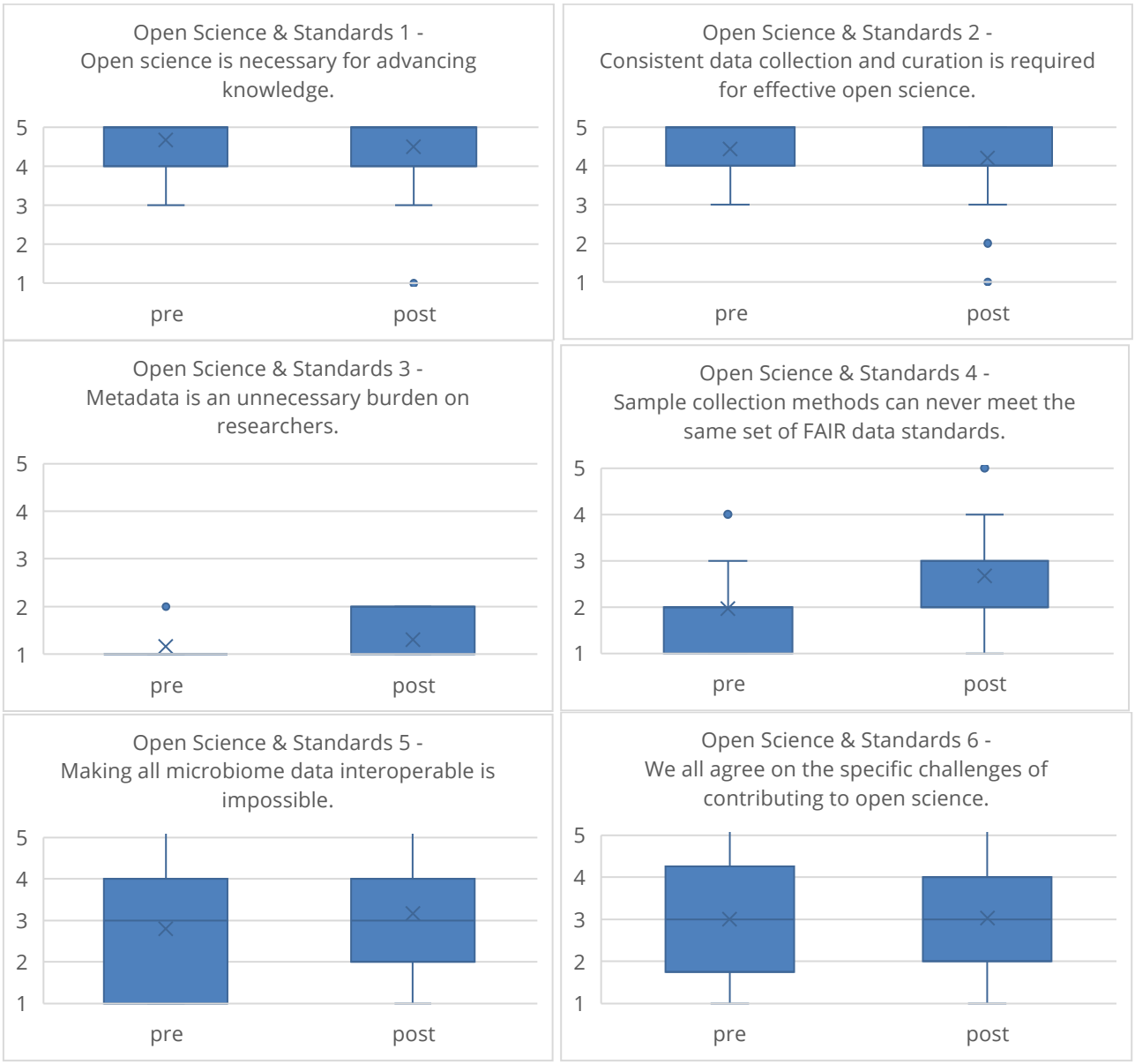


Fig. 1. Box and whisker plots for pre/post scores for the module "Open Science & Standards" (n=30)

Scale 1=disagree, 5=agree

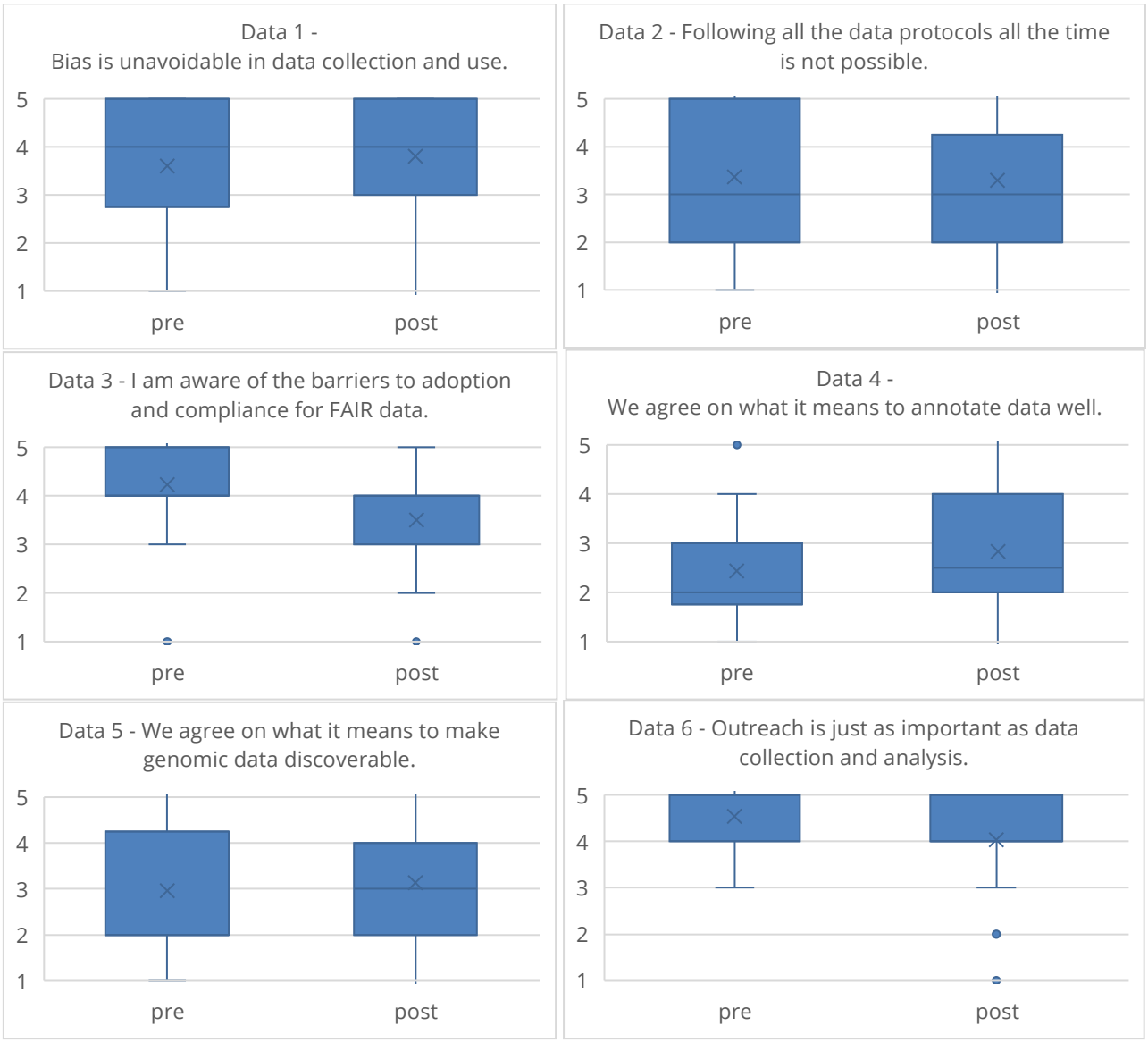


Fig. 2. Box and whisker plots for pre/post scores for the module "Data" (n=30)

Scale 1=disagree, 5=agree

X represents the median and the line through the box represents the mean

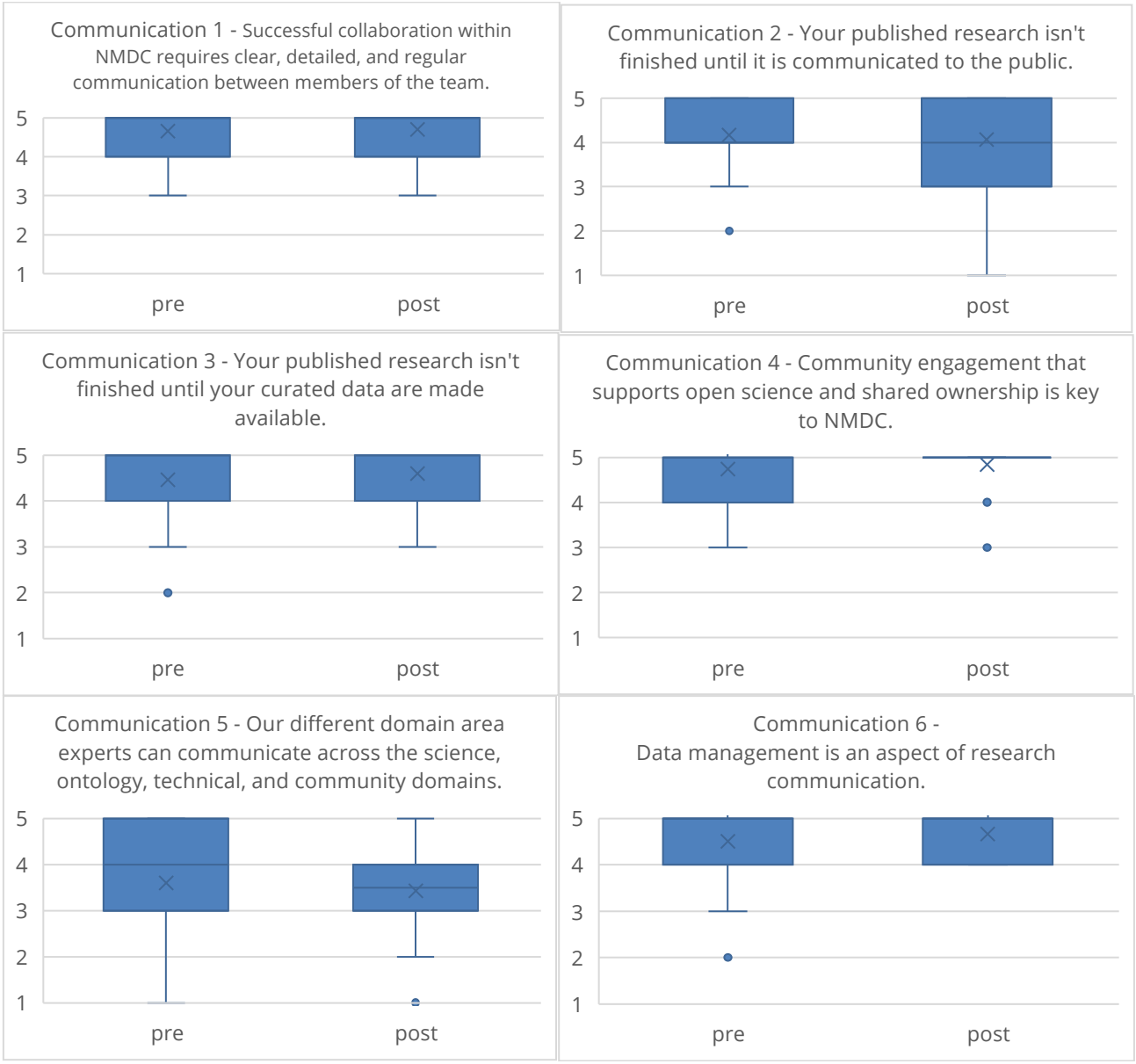


Fig. 3. Box and whisker plots for pre/post scores for the module "Communication" (n=30)

Scale 1=disagree, 5=agree

X represents the median and the line through the box represents the mean

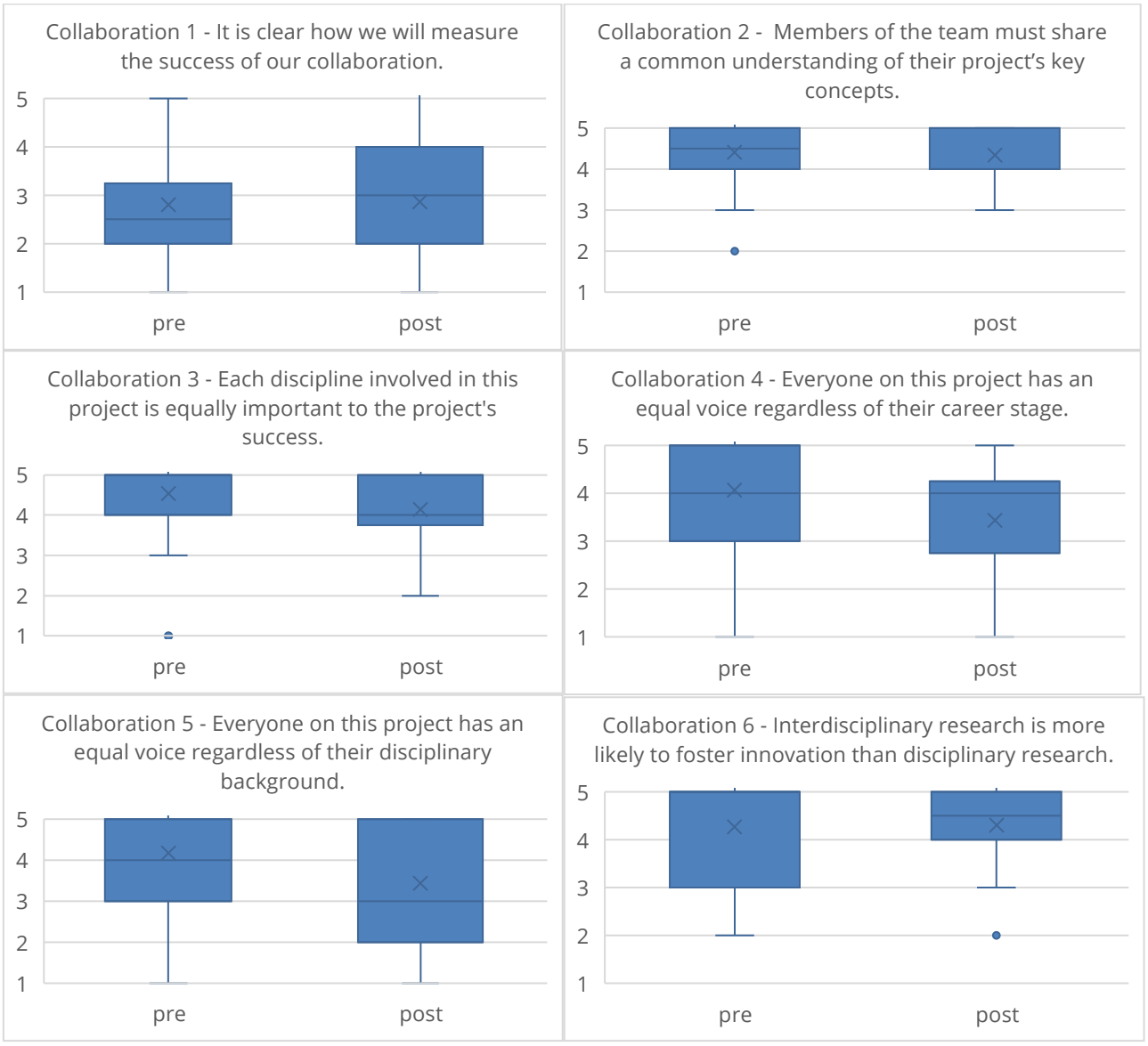


Fig. 4. Box and whisker plots for pre/post scores for the module "Collaboration" (n=30)

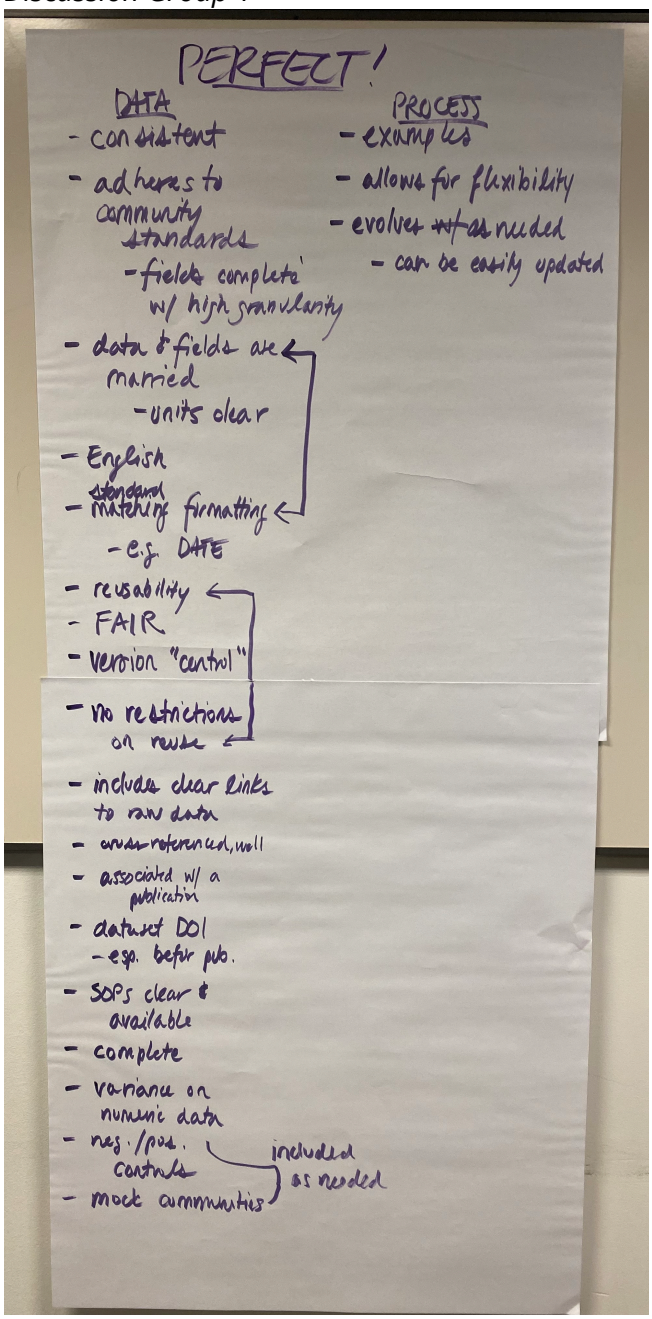
Scale 1=disagree, 5=agree

X represents the median and the line through the box represents the mean

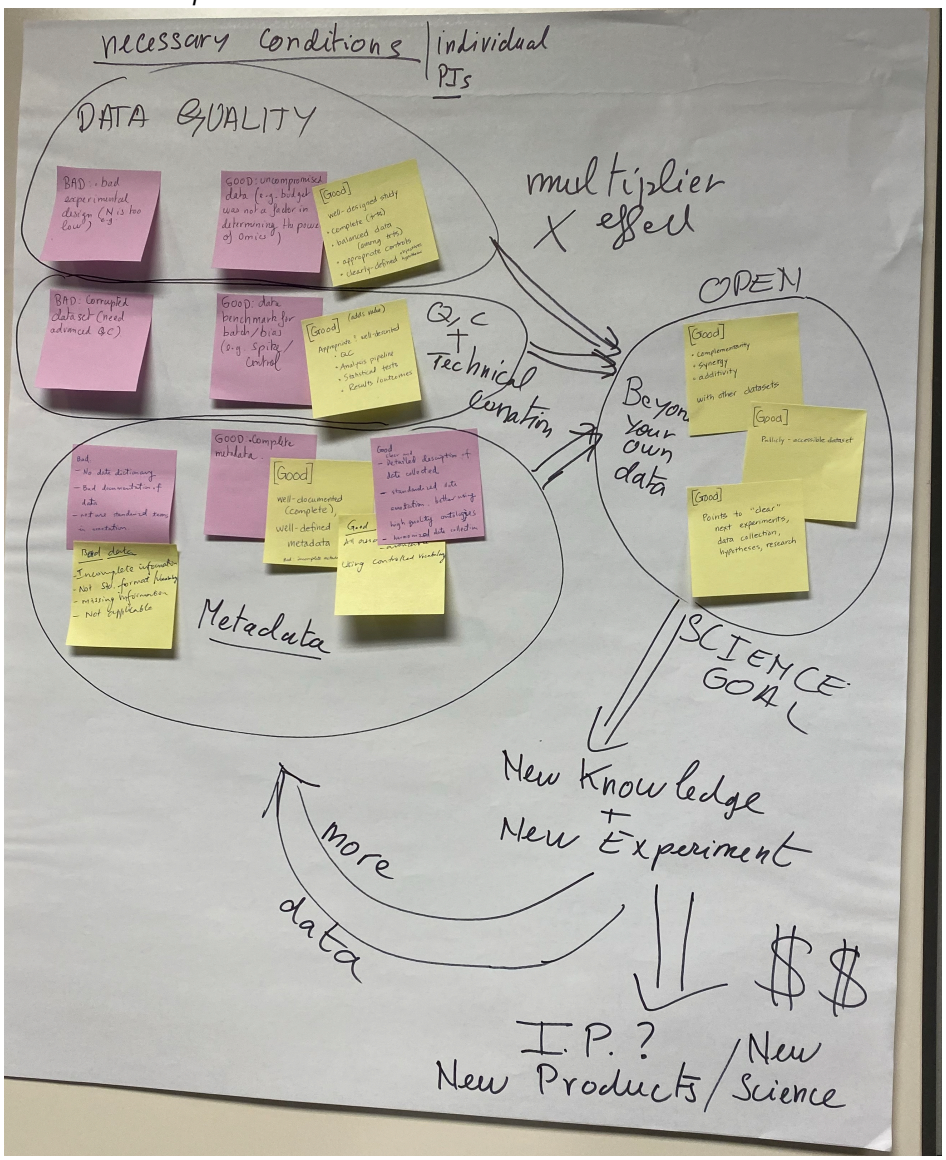
Appendix 3 – Co-Creative/Synthesis Activity

This appendix captures the thoughts from the final co-creation activity on ideal data sets.

Discussion Group 1



Discussion Group 2



Discussion Group 3