Cell Genomics Transparent Peer Review Record

Empirical Validation of an Automated Approach to Data Use Oversight

Moran N. Cabili[†,1], Jonathan Lawson [†,1], Andrea Saltzman[1], Greg Rushton[1], Pearl O'Rourke[3], John Wilbanks[4], Laura Lyman Rodriguez [5], Tommi Nyronen[6], Mélanie Courtot [7], Stacey Donnelly[1], Anthony A. Philippakis [*,1]

Summary

Scientific Editor:	Orli Bahcall
Initial submission:	2/28/2021
Revision received:	6/30/2021
Accepted:	8/07/2021
Rounds of review:	2
Number of reviewers:	3

Referee reports, first round of review

Reviewer #1: The article is a companion to the one presenting the Data Use Ontology (DUO) and reports on the DUO's evaluation in-use for the automation of data use agreements for data access. This a paradigm shift and a potential gamechanger for data access agreement management, a point of pain for biomedical researchers across disciplines and topics and one that will open up the possibility to dramatically accelerate data reuse and discovery science. From a technical perspective, the algorithm for matching data use requests to allowed usage conditions is an interesting application of ontology-driven engineering in practice.

My comments are just minor requests for clarifications.

The sentence "The DUO defines data use terms for secondary use, which are themselves imported from other ontologies, including the Monarch Disease Ontology (Mondo) (14), commercial use, populations (ethnicities and gender), methods development, and aggregate statistics." on page 7 is difficult to interpret. I first read this to be saying that commercial use, populations, methods etc. were all other ontologies alongside Mondo, then I realised that these were intended as examples of data use terms for secondary use. It is still not 100% clear to me though. Consider rewording or expanding slightly?

The reference 15 is where the main details of the matching algorithm is set out (and the source code is available). However, this reference is titled "The DataBiosphere consent ontology". This ontology is not mentioned in the paper nor in the companion DUO paper. This might be a bit confusing (it was for me). Is it worth including a brief clarifying (historical, I presume) note about this?

Reviewer #2:
General impression.

The paper describes a well known drawback in the access to data controlled by Data Use Agreements. The proposed solution of automating and standardizing provides a useful framework for addressing the bottleneck caused by an increasing number of both controlled datasets and of researchers applying to use them. They do not, however, discuss how to implement such a system. Who will evaluate and codify all the DULs for the thousands of sets of existing data? Perhaps suggest that any Data Access Request application come with its requirements already codified -- the requester is the one most motivated. That is half the process. However, what motivates those who grant access? Hire an army of people to read the existing DULs for thousands of datasets? And do it twice or thrice to ensure they match? Or maybe coding each dataset as it is requested? That might gradually build up a group of datasets ready for the automated comparison, but in the meantime it slows down the process for each application. Perhaps the low number of examples (fewer than 100 for each test in Figure 3) speaks to the difficulty of implementation.

A useful approach for the future might be for the data repositories to require that new data come pre-annotated wrt DULs. Standardized DULs might be useful to reduce the large number of different types and versions. How would that be implemented? What are the obstacles -- what comes to mind is the difficulty of getting the Institutional Review Boards of all the hospitals and universities to converge on a few protocols when each is looking over its shoulder at a bank of lawyers lining up to get them.

Specific issues.

abstract: "non-inferior" to other methods to evaluate match btw DUL and DAR? If they believe the system is only non-inferior, then they are not making a very strong argument for it.

Fig 2. One wonders if the data are from the same input. The totals seem too smooth. Y-axis on 2b says "datasets released per quarter". Is that correct? seems not.

Fig 3. Would be nice to see more detail in the right side of the images. Perhaps expand the width of the image to full-width and put the text on the left side above each segment. It might be useful to have the number of items that were evaluated in each part of the figure mentioned in the figure. 100% of what? 5? 500? 10000? The numbers are in the text (and rather low), but should be in the figure legend as well.

p.6
"Moreover, we describe the creation of a production-grade software platform called "DUOS" that implements the DUO standard,
automates data use oversight and simplifies the DAC workflow even when traditional nonautomated approaches are used."

But they do not discuss how exactly it might simplify the DAC workflow for traditional nonautomated approaches. Perhaps it is obvious, but some mention of standardized output from the DUOS might be appropriate.

"Prospectively, the GA4GH Machine Readable Consent Guidance (11) seeks to reduce the frequency of consent forms that cannot be mapped to DUO by offering directions and template consent form language."

good idea. I was thinking the same thing. How to get IRBs to adopt them?

It would be a good idea to see some examples of some DUO ontology.

p.7
see companion manuscript by Voisin. what paper is that?

Grammar nit.

p.6.
Finally, we note that, in all cases (51 out of 51) that were reviewed by the DAC and DUOS, both DAC and DUOS were in agreement

drop "both". As written, it makes one wonder what third thing they both are in agreement with.

Reviewer #3: Comments enter in this field will be shared with the author; your identity will remain anonymous.

Thank you for the opportunity to review this paper. It presents an evaluation of an automated approach to data access approvals which makes use of GA4GH standards -- these demonstrations are crucial in driving broader implementation. The manuscript and figures are very clear, and I don't have any suggestions for improvement.

Author response to the first round of review

Editor's comments:

We have now received the final reviewer report on your paper, and a copy of all 3 reports are attached below. We invite you to revise your manuscript in response to these referee comments and the editorial requests below. I will be glad to discuss how to best focus and present this work, in context of this manuscript, and as part of the special issue.

Author response:

We are grateful for the Editor's and Reviewers' thoughtful comments and questions, and for the opportunity to revise this manuscript. We believe that the Reviewers have raised a number of important points, and we have sought to address these points in the enclosed point-by-point response, as well as in revisions to our manuscript. We feel that our revised manuscript has been substantially improved as a result of the Editor's and Reviewers' inputs.

We address these reviews in more detail below.

1) Please revise to present the manuscript in our Technology article format. (2nd tech format reference)

Manuscript change:

We have changed our manuscript to fit the "Technology" article format. In particular, we have

1) added a "Design" section that outlines the challenges to our current framework for data access, as well as an analysis of data from dbGaP providing evidence that these challenges are real.
2) added a "Limitations" section to the Discussion section.

2) Please include a STAR Methods section (see details below).

Manuscript change:

We retitled the current Methods section to "STAR Methods" and added "Lead Contact," "Materials Availability," "Data and Code Availability," and "Methods Details" subheadings with appropriate content.

3) You may include additional supporting information in Supplementary Information (see details in email).

Author Response:

We now included a supplementary methods section that provides a more detailed overview of the DUOS matching algorithm and interfaces.

4) Please also include a Highlights and eToc and Graphical Abstract with your revision. These should describe the context and significance of the work for a broad readership. The goal is to highlight the major advances in the paper in order to attract the attention of the non-specialist, without including extensive detail.

Manuscript change:

Thank you for the comment. We have now added Highlights, eToc, and a Graphical Abstract as suggested.

5) Methodological details: It appears that essential methodological details were not yet included in this manuscript. As this is the first primary publication of DUOS, please include all methodological details are included in this manuscript main text, STAR Methods and/or Supplementary Information. Please also ensure that complete methodological details for all analyses reported in the manuscript are included.

Author response:

Thank you for the suggestion. We now added sections in STAR Methods and Supplementary Methods that provide an overview on how the DUOS algorithm matches data access requests with dataset use restrictions, and how the system encodes these elements using an ontology. We provide specific examples as well as tables that specify the encoding rules and matching rules.

If any further clarifications are required, we are happy to provide them.

6) Please introduce all concepts at first use, for a general readership. For example: for library card, please introduce the concept, how this works in practice, and your implementation.

Manuscript change:

Thank you for the suggestion. We have now updated the text in p. 7, to clarify two key issues that will make them more accessible to a general audience

- The concept of GA4GH passports
- The concept of a library card.

If these explanations are still too technical, we are happy to revisit. Also, if there are other "terms of art" related to data governance that are still too technical for a general audience, we are happy to revise (we have read through the manuscript with an eye towards this and did not see any; however the authors may be a bit too close to the nuances of data governance to appreciate what may not be obvious).

7) For in progress or planned studies mentioned (such as pilot projects with NIH): please provide full details to explain the current or planned projects and their status.

Author response:

Thank you for the suggestion, we have now updated the text in p. 6-7 to more fully describe the pilot with NIH.

8) Please include a Limitations sub-section in the discussion.

Author response:

We have added it.

9) Please include a forward looking section at the end: discuss your future plans for developing DUOS and other related methods, continuing pilots and broader implementations.

Author response:

We have added a “Future Directions” section at the end of the “Discussion.”

We hope that you will be able to suitably revise within the next 3 weeks. If you will need additional time, please keep us updated on your progress. I will be delighted to discuss these and additional editing suggestions and your plans for revising, to support you in the optimal presentation for this exciting and impactful publication.

Author response:

We are happy to discuss any further ways of improving this manuscript.

Reviewer #1

Reviewer #1: The article is a companion to the one presenting the Data Use Ontology (DUO) and reports on the DUO's evaluation in-use for the automation of data use agreements for data access. This a paradigm shift and a potential gamechanger for data access agreement management, a point of pain for biomedical researchers across disciplines and topics and one that will open up the possibility to dramatically accelerate data reuse and discovery science. From a technical perspective, the algorithm for matching data use requests to allowed usage conditions is an interesting application of ontology-driven engineering in practice.

My comments are just minor requests for clarifications.

The sentence "The DUO defines data use terms for secondary use, which are themselves imported from other ontologies, including the Monarch Disease Ontology (Mondo) (14), commercial use, populations (ethnicities and gender), methods development, and aggregate statistics." on page 7 is difficult to interpret. I first read this to be saying that commercial use, populations, methods etc. were all other ontologies alongside Mondo, then I realised that these were intended as examples of data use terms for secondary use. It is still not 100% clear to me though. Consider rewording or expanding slightly?

Manuscript change:

Thank you for this helpful suggestion. We have reworded this as follows as part of the STAR Methods section on p.14:

"The DUO defines terms for secondary data use, such as general research use, commercial use, populations research (ethnicities and gender), methods development, aggregate statistics as well as disease-specific research which relies on terms imported from the Human Disease Ontology (Schriml et al. 2018). The details of DUO are described in the companion manuscript, but it can most simply be thought of as a cross-product of sub-ontologies, one for each of the above fields."

The reference 15 is where the main details of the matching algorithm is set out (and the source code is available). However, this reference is titled "The DataBiosphere consent ontology". This ontology is not mentioned in the paper nor in the companion DUO paper. This might be a bit confusing (it was for me). Is it worth including a brief clarifying (historical, I presume) note about this?

Manuscript change:

Per editor guidelines we moved this text to the Data and Code Availability section, and added supporting text per the reviewer comments:

"The code for DUOS is distributed under the three-clause BSD open source license in the "DataBiosphere" GitHub repositories located here:

- https://github.com/DataBiosphere/duos-ui
- https://github.com/DataBiosphere/consent
- https://github.com/DataBiosphere/consent-ontology
- https://github.com/DataBiosphere/consent-data-use

The "DUOS-UI" is the front end user interface for the application. "Consent" is the application component where all user, data access requests, and dataset access decisions are stored. "Consent Ontology" is the application component that can make ontologically valid decisions about whether a dataset and a data access request are compatible using the GA4GH (DUO) and Human Disease Ontology (DOID) ontologies. "Consent Data Use" captures all the ontologies we referenced in the article for efficient access by the DUOS service.

In addition, we have made DUOS available as a publicly accessible web-service at https://duos.broadinstitute.org.

Data reported in this paper will be shared by the lead contact upon request.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request."

We have also added an extended description of the details of the DUOS matching algorithm in STAR Methods and Supplementary Methods sections.

Reviewer #2

General impression.

The paper describes a well known drawback in the access to data controlled by Data Use Agreements. The proposed solution of automating and standardizing provides a useful framework for addressing the bottleneck caused by an increasing number of both controlled datasets and of researchers applying to use them.

They do not, however, discuss how to implement such a system. Who will evaluate and codify all the DULs for the thousands of sets of existing data?...However, what motivates those who grant access? Hire an army of people to read the existing DULs for thousands of datasets? And do it twice or thrice to ensure they match? Or maybe coding each dataset as it is requested? That might gradually build up a group of datasets ready for the automated comparison, but in the meantime it slows down the process for each application.

Author response:

Thank you for raising these important concerns. We very much appreciate the prompt to highlight the operational challenges ahead in seeing DUO and DUOS move beyond this initial pilot. We have added substantial text to the manuscript to reflect the importance of this point, and feel this input has greatly strengthened the manuscript. Thank you!

We describe these changes in the several answers below, each corresponding to one of the Reviewers' comments. In this first one, regarding structuring of DULs,there are two distinct answers, depending on whether we are prospectively structuring DULS as part of new studies, or retrospective structuring DULs of legacy studies.

For prospectively structuring DULs of new studies, the answer is more straightforward. In addition to the DUO standard, GA4GH has also created the Machine Readable Consent Guidance (MRCG). This provides guidance for human subjects researchers and IRBs drafting consent language so that it maps onto DUO terms. Assuming it gains traction (see below), then mapping DULs to the DUO standard will be straightforward for new cohorts.

For retrospective cohorts, the reviewer raises an important point. We feel that the following considerations are important to note on why it is realistic to think that data custodians will seek to structure their DULs:

1) It takes only a small amount of time for a person with knowledge of DUO to map a DUL to it. For example, in a recent effort, a single person was able to map over 600 DULs to DUO in a small period of time (weeks, but doing so in a far from full-time capacity).
2) DACs field a large number of requests for access to data where the research purpose is clearly at odds with the data use restrictions (e.g., a request for studying diabetes on a dataset that is only consented for cancer); dbGaP tells us that the vast majority of data access requests that they decline are for obvious misunderstandings of the DULs. This is a significant waste of the DAC's time. By allowing a researcher to automatically "filter" to the datasets that are consistent with their research purpose, these erroneous requests are obviated.
3) It is important to note that each DUL needs to be structured only once. An essentially infinite number of DARs can then be matched against it once it is structured. So the work is linear in number of datasets, but exponential in the benefits to both researchers and DACs in the time saved by streamlining (and eventually automating) the process of data governance. This is a powerful motivation.

Consistent with this, we find that groups managing large numbers of datasets are quite open to structuring their DULs with DUO. In addition to the Broad experience described in this study, other efforts underway include

- As stated in the manuscript, we are now undertaking a large-scale effort across 7 NIH Institutes and Centers (ICs) to structure their DULs using DUO. Over 600 of the ~750 that these ICs govern have been structured using DUO.
- As of March 2019, the European Genome Phenome Archive (EGA) asks all data submitters to align their data access policies with DUO so that their dataset(s) can be tagged appropriately (see https://ega-archive.org/blog/data-use-ontology/)

Manuscript change:

We added the following text in p.6 to the manuscript to address these important questions raised by the reviewer:

"First, there is the task of retrospectively coding the many existing DULs to the DUO standard. Here, we are encouraged by our initial experiences with the large-scale NIH trial mentioned above. There, a single individual knowledgeable in the DUO standard was able to codify over 600 NIH datasets in a matter of weeks, working far from full time. It is our experience that DACs are often highly motivated to code their DULs, as it can result in many fewer requests for access to data where the research purpose is obviously at odds with the DUL (certainly, this is true in the dbGaP experience), lessening unproductive work of the DAC. Moreover, the DAC at our institution now requests data depositors to send the DUL in a template form that asks simple questions like, "Are there restrictions on commercial use?" and "Are there any disease-specific restrictions?" that correspond to DUO fields. These are easy for an investigator or signing official to fill out and greatly simplifies the task of mapping the DUL to DUO. We believe that it is feasible for other DACs to start adopting a similar approach to lightweight formatting of the DULs they receive.

Second, there is the task of changing our approach to prospectively consenting patients. Here, IRBs and PIs will be greatly aided by the GA4GH Machine Readable Consent Guidance (MRCG), which ensures consent forms include data use language in verbatim DUO terms and definitions, through the template language the MRCG provides. A significant task will be convincing IRBs and researchers designing consent forms to adopt this guidance. Achieving this will require both evangelizing the benefits of DUO, as well as providing empirical evidence of their validity and utility in real world data governance. We hope that this work represents a first step towards providing such an evidence base."

Perhaps suggest that any Data Access Request application come with its requirements already codified -- the requester is the one most motivated. That is half the process.

Author response:

We certainly agree with the reviewer's perspective. Currently in DUOS, researchers submitting Data Access Request applications are required to codify their applications with DUO terms which accurately describe their research purpose. These are then subsequently validated by the Data Access Committee who reviews both the written and codified research purpose.

It is an interesting idea to additionally ask data requestors to share the load of structuring DULs. However, one could argue that they may be incentivized to do so in a manner that interprets them more leniently than intended. For example, a dataset whose DUL states "Diabetes and related conditions" would be structured as "Diabetes only" in our current approach, under the rationale that it is better to be conservative and not allow inappropriate use of data. A researcher requesting access to this data might take a much more liberal view of what is appropriate secondary use.

Nevertheless, we agree that it is an interesting idea and will strongly consider experiments with it in future efforts.

Perhaps the low number of examples (fewer than 100 for each test in Figure 3) speaks to the difficulty of implementation.

Author response:

Thank you for highlighting this concern. The low number of examples is actually due to intentionally limiting the number of datasets permitted in the pilot under the IRB protocol that authorized the DUOS trial - under 10 for most of its course. We are now processing large numbers of DARs as part of the NIH pipot described above.

A useful approach for the future might be for the data repositories to require that new data come pre-annotated wrt DULs. Standardized DULs might be useful to reduce the large number of different types and versions. How would that be implemented? What are the obstacles -- what comes to mind is the difficulty of getting the Institutional Review Boards of all the hospitals and universities to converge on a few protocols when each is looking over its shoulder at a bank of lawyers lining up to get them.

Author response:

We strongly agree with this suggestion, and the Broad has recently adopted it in our approach to receiving DULs. While in the past, we allowed data depositor to send the DUL in free text, we know send them a template that structures the way they send the DUL by asking questions like:

- Are their commercial restrictions?
- Are their disease specific restrictions?
- Is the data available for methods development.

Using this template has made it markedly easier to map the DUL to DUO, and is not at all laborious for the data depositor. It is a lightweight approach that does not require IRBs to converge on some standard protocol, nor does it require legal review. It merely asks the data depositor to check boxes in a structured form.

We also feel that this approach will propagate to other DACs. Currently, NIH requires only that the Signing Official for each institution depositing data to sign-off on the data use limitations, while use of IRB or privacy board to aid in determining is referenced but not required. The European Genome-Phenome has a similar practice in which they ask all data "submitters to align their data access policies with DUO so that their dataset(s) can be tagged appropriately".

Manuscript change:

We have now highlighted this this point on p7 to highlight the important point that the reviewer raises:

"First, there is the task of retrospectively coding the many existing DULs to the DUO standard. Here, we are encouraged by our initial experiences with the large-scale NIH trial mentioned above. There, a single individual knowledgeable in the DUO standard was able to codify over 600 NIH datasets in a matter of weeks, working far from full time. It is our experience that DACs are often highly motivated to code their DULs, as it can result in many fewer requests for access to data where the research purpose is obviously at odds with the DUL (certainly, this is true in the dbGaP experience), lessening unproductive work of the DAC. Moreover, the DAC at our institution now requests data depositors to send the DUL in a template form that asks simple questions like, "Are there restrictions on commercial use?" and "Are there any disease-specific restrictions?" that correspond to DUO fields. These are easy for an investigator or signing official to fill out and greatly simplifies the task of mapping the DUL to DUO. We believe that it is feasible for other DACs to start adopting a similar approach to lightweight formatting of the DULs they receive."

Specific issues.

abstract: "non-inferior" to other methods to evaluate match btw DUL and DAR? If they believe the system is only non-inferior, then they are not making a very strong argument for it.

Author response:

Thank you for this suggestion. For context, we have attempted to run this study as a true trial by blinding the DAC members to DUOS and each other. We chose the phrase, 'non-inferior' as it is commonly used in the trial literature when attempting to show that two interventions are equivalent, up to the statistical power of the trial. However, we recognize that this may be a bit too subtle and confuse some readers. Therefore, we have adopted your suggestion.

Again, we would like to stress that we are comparing DUOS to a 'gold standard' of 5 experts in data privacy and human research subjects protection sitting on our DAC. *We would not imagine DUOS to be more accurate* than them at adjudicating data use oversight; rather an approach that is *comparable to them but automated would be transformative* in accelerating access to data from human subjects.

Manuscript change

We have replaced 'non-inferior' with 'comparable'.

Fig 2. One wonders if the data are from the same input. The totals seem too smooth. Y-axis on 2b says "datasets released per quarter". Is that correct? seems not.

Author response:

Thank you for noting this. We have just noticed that the original figure 2b was labeled with “datasets released per quarter” when it should have been labeled as “cumulative datasets.”

We have now fixed this. Also, we now display the panels on the left side using bar plots, and we have changed the group size to year, rather than quarter, to more clearly present the incremental number of new datasets (2a), requesters (2c), and DARs (2e) each year and the corresponding cumulative distribution on the right panels, (2b), (2d) and (2f), respectively. We now included the ***Quantification and Statistical Analysis*** section in STAR Methods that describes the details of this analysis.

Regarding the smoothness of the plots on the right, we have checked and it is accurate.

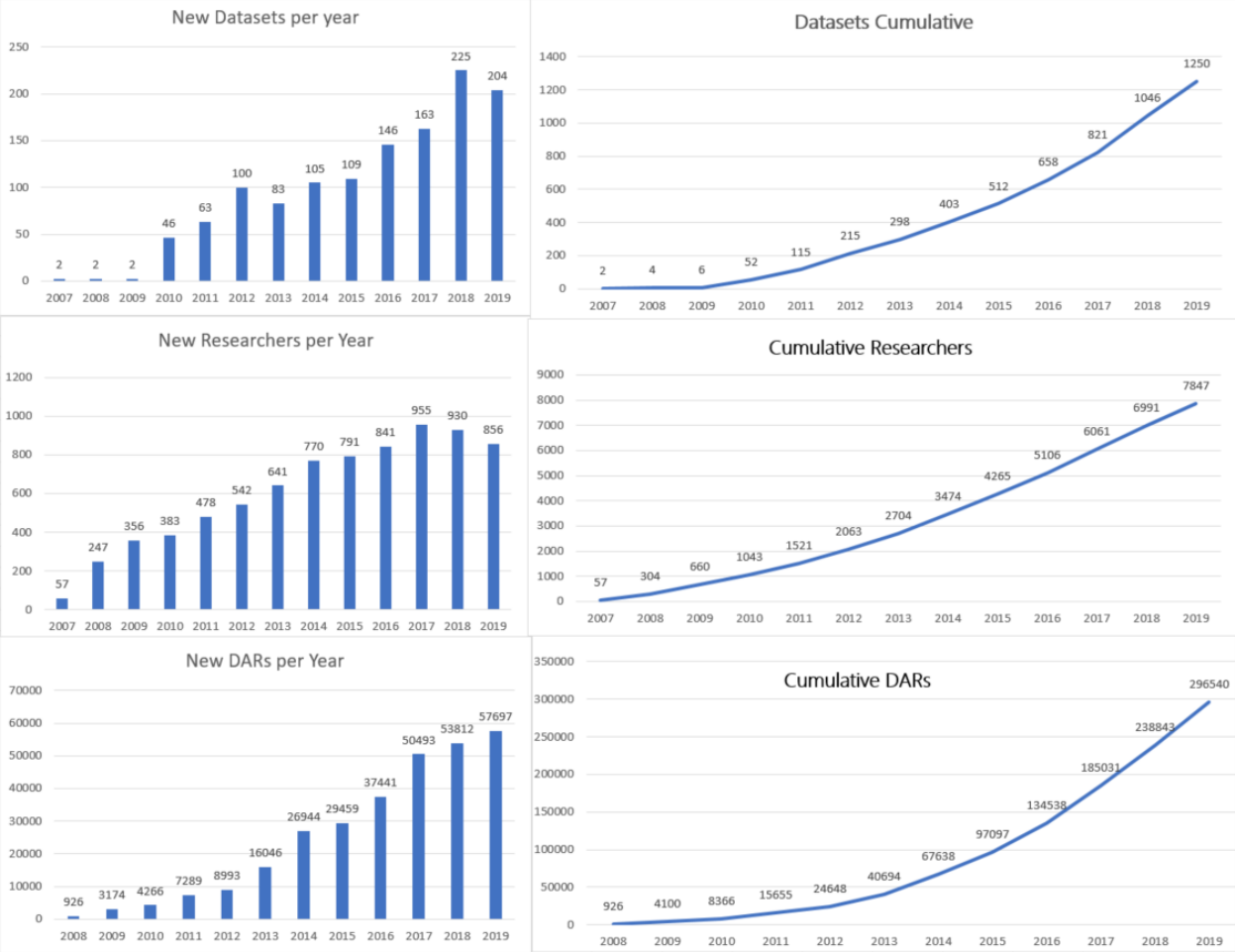

Fig 3. Would be nice to see more detail in the right side of the images. Perhaps expand the width of the image to full-width and put the text on the left side above each segment. It might be useful to have the number of items that were evaluated in each part of the figure mentioned in the figure. 100% of what? 5? 500? 10000? The numbers are in the text (and rather low), but should be in the figure legend as well.

Manuscript change:

We thank the reviewer for this suggestion and agree with it. We have modified Figure 3 by expanding to full-width, and added the N to each panel.

p.6

"Moreover, we describe the creation of a production-grade software platform called "DUOS" that implements the DUO standard, automates data use oversight and simplifies the DAC workflow even when traditional nonautomated approaches are used."

But they do not discuss how exactly it might simplify the DAC workflow for traditional nonautomated approaches. Perhaps it is obvious, but some mention of standardized output from the DUOS might be appropriate.

Author response:

We thank the reviewer for highlighting this point and agree that it is worth providing additional text on this point.

***Manuscript change*:**

We have added the following text on p5:

"We call the software package that encapsulates these three functionalities "DUOS" (short for "Data Use Oversight System"). Moreover, knowing that there will always be datasets whose DULs cannot be structured, DUOS also provides user interfaces to support the workflows of traditional DACs. The DUOS user interfaces substantially facilitate the work of a DAC, even in the non-automated setting (which is the vast majority of cases today), by: 1) clearly presenting the DULs and DARs to the DAC, and simplifying how they track progress by compiling the number of DAC members that have voted and the results of their votes, 2) sending notices to DAC members when new DARs come in or when they are delayed in voting, 3) demonstrating to DAC members if researchers have been approved by their institution to submit DARs, 4) allowing researchers to apply for access to many datasets from multiple DACs through a single DAR application form, thus lessening the number of DARs that need to be submitted, 5) providing an API that enumerates the "white-listed" researchers that have been approved to access various dataset, removing the need for communication between the DAC and the growing number of data repositories hosting these datasets. Support for DACs is an enormous unmet need that we hope that systems like DUOS and related systems like AAI REMS, a Resource Entitlement Management System developed by Elixir Finland, can fill (Linden et al., 2018)."

"Prospectively, the GA4GH Machine Readable Consent Guidance (11) seeks to reduce the frequency of consent forms that cannot be mapped to DUO by offering directions and template consent form language." good idea. I was thinking the same thing. How to get IRBs to adopt them?

It would be a good idea to see some examples of some DUO ontology.

Author response:

We thank the reviewer for this suggestion, and we think it is worth clarifying that there is a companion manuscript that describes the creation of the DUO standard in detail. It contains substantial documentation of how DUO works (for example, we are re-pasting a picture of it below), along with examples of DULs that have been structured with DUO.

We agree that it would be good to provide examples but feel that it is better to do this by pointing to the companion manuscript. We have made reference to this in the main text (below).

It is a good point about adoption of the Machine Readable Consent Guidance document. We argue that the best path is by providing empirical evidence of the validity and utility of DUO in streamlining data oversight, which this manuscript is a real attempt to do. We have added text to point to the importance of adoption.

Manuscript change:

On p4-5, we address the point on examples with the text "The resulting Data Use Ontology (DUO) is now a GA4GH standard. (see companion manuscript, which provides extensive detail on how it is structured and examples of its use in practice)"

On p 8, "Second, there is the task of changing our approach to prospectively consenting patients. Here, IRBs and PIs will be greatly aided by the GA4GH Machine Readable Consent Guidance (MRCG), which ensures consent forms include data use language in verbatim DUO terms and definitions, through the template language the MRCG provides. A significant task will be convincing IRBs and researchers designing consent forms to adopt this guidance. Achieving this will require both evangelizing the benefits of DUO, as well as providing empirical evidence of their validity and utility in real world data governance. We hope that this work represents a first step towards providing such an evidence base."

p.7

see companion manuscript by Voisin. what paper is that?

Author response:

As stated above, this is a complementary paper co-submitted by GA4GH colleagues with this manuscript describing the passport.

Grammar nit.

p.6.

Finally, we note that, in all cases (51 out of 51) that were reviewed by the DAC and DUOS, both DAC and DUOS were in agreement drop "both". As written, it makes one wonder what third thing they both are in agreement with.

Manuscript change:

Replaced 'both' with 'the'.

Referee reports, second round of review

Reviewer #1: The authors have addressed all the points I raised before, thus I am pleased to recommend the work for publication.

Reviewer #2: Authors have addressed issues raised by the reviewers.

Reviewer #3:
The authors have satisfactorily addressed the reviewers' feedback.

Author response to the second round of review

Thank you very much for preliminarily accepting our manuscript for publication, and for the thoughtful comments and suggested edits! We have addressed your requests, as stated in bold below:

1. *Please see the attached files, which include our detailed edits to the text, all noted with tracked changes and/or Comments. These edits are intended to improve the clarity, presentation and reporting in the manuscript.* **Thank you for the edits. We have accepted these edits and have**

added additional edits to address your comments in track changes. Please see the attached track changes draft.

2. *We will be publishing as a Technology article. Thank you for formatting accordingly*. **We believe the article is now in the proper format. Please let us know if any further action is required.**
3. *Please include a cover letter that details all changes made in the revised manuscript (you may also include an additional manuscript file with tracked changes).* **Please see the version updated with track changes in which we addressed your comments. In summary:**
    a. **We added a brief intro on dbGaP.**
    b. **We added a reference to the DUO paper that explains how "how DUO handles reviewing and updating".**
    c. **We added a description of the datasets analyzed and a time frame for the analyzed requests.**
    d. **We added citations to the GA4GH marker, DUO, and GA4GH Passports papers.**
    e. **We added more detail on current challenges around data access.**
    f. **We added more detail on the previous and current DUOS pilot activities with NIH.**
4. *Limitations - thank you for including this section in the Discussion. The goal of this section is to promote clarity and transparency by highlighting any limitations in the interpretation of the study, including limits of the techniques used and/or assumptions made. It can include additional experiments that would be necessary to definitively prove some conclusions but should be specific to the paper. Examples include sample size, genetic strains, detection levels, etc. The "Limitations" paragraph should be specific to this paper.* **Thank you. We included your suggested edits per our correspondence.**
5. *Please include a 'Highlights and eToc' and 'Graphical Abstract' with your revision. These should describe the context and significance of the work for a broad readership. The goal is to highlight the major advances in the paper in order to attract the attention of the non-specialist, without including extensive detail. Highlights are 3–4 bullet points of no more than 85 characters in length, including spaces, and they summarize the core results of the paper in order to allow readers to quickly gain an understanding of the main take-home messages. The eTOC blurb should be no longer than 50 words and describe the context and significance of the findings for the broader journal readership.* **This was previously submitted, and is reattached to this email.**