**FINDING THE INFORMATION THAT SCIENTISTS NEED**

**1. What types of information might be “linked” to preprints to facilitate interpretation of data?**

 -Link preprint to final paper and vice versa

 -Incorporate feedback mechanisms on preprint article; such as blogsites, Twittering and direct commentary on papers

 -Incorporate easy ways for authors and readers to interact

 -Find ways to “map” these interactions into peer-review system

 -Make preprint visible to Google Scholar, Pubmed, Europe PMC, WoS, Scopus

 -flag papers to indicate “not yet peer reviewed” status

 -papers should have assigned DOI and be hygiene checked

 -Image data and computer code in papers should be made available, if possible.

 -encourage and create more value in submitting large data sets

 *a) How can this additional information be managed so scientist can shift through preprints and find relevant information?*

 Need to move to one biology server to ensure that Google searches can run smoothly through the data. Start simply, with the goal of transitioning to a more structured submission process and paper format. Authors should all have ORCHID accounts to facilitate searching. They should include key words and categories for their work. Papers should use searchable form of abstract and title. Arrange for all papers to have standard appearance, with structured information and figures (not just blobs of text with figures hard to find). Provide easy way to access by highlighting key words and concepts. Want the format to be organized in such a way that the system can embrace a new complex mode of presenting data.

**2. What improvements can be made to preprints?**

 -Allow large data sets to be associated with the preprint, as well as videos and novel organizational tools.

 -Ensure a mechanism for checking for data manipulation. This could be a real problem as authors become anxious to get their data out.

 -Have links to large data sets.

 -Have 3rd party data base for commentary overlay.

 -Use computer-readable format for papers (not pdf).

 -Make new ‘cool’ tools for reviewing and archiving

 -Have mechanism for adding new data to an old preprint.

 -With big data sets, allow readers to easily contact and download data; have way for readers to interact with data and potentially report new findings from the preprint data.

 -Support browser so it can sort by topic/subject

 -Try out new experiments; put up 3-D reconstructions of neural networks or 3-D lattice light sheet movies of specialized processes for readers to mine the data themselves.

**3. What kinds of tools and evaluation methods can and should be used to help readers search for material by topic, date and investigator; as well as stratify preprints by quality, viewership, significance, novelty, etc.**

-Use bibliometrics (e.g., measure the #downloads per article, # citations)

**4. Other issues:**

 -What will be the commercial aspect of the preprint? Need the Preprint service to be managed by a trusted party. Non-profit organization should probably be running the service.

 -Bring preprint servers into one entity, since silos could be a problem. If the Preprint system is one entity, the search problem is solved. If many entities in a federation, who pays and how can we stop search engines from faltering?

 -May want a consortium of philanthropy to cover the costs of the enterprise.

 -Should have Genbank submission at time of preprint submission (same precondition as journals)

**From Ralf Bundschuh**

**Report on breakout group A on topic "Finding Information"**

**=========================================================**

> Note: lines with caret ">" have been added after the fact based

> on my memory of the discussion. Lines without caret are

> as taken during the discussion.

> First question to be considered:

What types of information should be linked to preprints?

> First resolution:

Critical to link preprint to final paper.

Journals require links to primary data, preprint databases should as well.

> At this point a somewhat lengthy discussion occurred that

> focused on the idea if preprint repositories should also be

> data repositories.

NSF: several large observatories, data repositories, would be useful for

preprints to link to existing data rather than putting copies on data

Many fields have standards for depositing data

What about data that does not fit current databases? E.g., videos.

Contributor's web site is not stable; journals should be responsible to keep

such data long term. Also differs by journal.

Volume in the life sciences is a problem. Is it a dollar problem?

Recent study: 12% of data associated with PMC papers is in recognized archives.

Where would TB data sets of raw data go?

Open data is different from preprints.

> Second resolution (if we do not want to get started on preprint until

> the larger problem of open data is solved, which not even the journals

> have really solved, we will simply not get to have preprints see the light

> any time soon):

Don't make preprints dependent on open data.

Example is zeka virus data and publication sharing agreed upon within 48 hours.

> Third resolution:

Use of ORCID very important.

> From here on a general discussion started on how to make preprint

> servers searchable. There was some arguing for a structured format

> of representing metadata as well as the article, but also arguing

> against it since the more input is required by users at submission

> time the higher the barrier for adoption. Consensus was that while

> more structured data would be nice, this should be slowly adopted and

> possibly be optional - once users see the advantage of providing more

> searchable, structured data, they might be more willing to invest the

> time of entering it.

Use of metadata/keywords/full text search can help.

Needs guide to browsing - categories.

Licensing is critical to reusability.

Needs computer readable formats, not just PDF.

Is there a need to impose more structure rather than just blob of text.

Or at least allow it and use a common meta-format.

> While I do not remember how we got into this, we did have a discussion

> about the one preprint server versus many question.

Will there be one preprint server or many?

Currently: bioRxiv, arXiv, PeerJ, PLoS, F1000

PDB (protein data bank) was able to come up with a governance model where

criteria for incorporation are shared. Spreads costs, same data for all,

in principle each site can have its own services, but some redundancy developed.

Tradeoff between creativity of ways to offer information and making things easy

for the user who wants only one place of information.

Community building is the main goal. Preprints are just a sliver of this but

should integrate with all kinds of data. This will allow other people to

mine the data. There could be many different

communities within the biomedical sciences.

Counter example: cibig(?) (NCI) tried to set up one database, but failed,

process was not good since not community driven.

> And back to the question of what to do with the data.

Culture change is required for scientists to be willing to share their data.

Storing large data sets is a cost issue.

Linkage of preprints and data is a problem.

How is data being credited? NIH is working on data citations. So in addition

to papers and maybe preprints there will also be data.

Who is enforcing data submission for preprints.

> Fourth resolution:

Where possible data associated with the preprint should be

submitted at appropriate community servers.

> This comment referred to the question of a data deposition server again:

Is there any form of intermediate step to really create more value than

currently the case?

> Fifth resolution:

Need another meeting like this on the issue of data deposition.

> Now, we had some discussion about the timing of data sharing versus

> preprint sharing.

Putting data out before paper is published might create even more of an issue

with the fear of being scooped.

Genetics community: cannot publish until the authors have published first

paper on this data. Question: does preprint comprise publication in this

context?

Is there a need for central data repository if investigators already keep their

own data and could just share it.

As far as preprints are concerned, it is important to keep it simple. Tethering

preprints to data storage could be kiss of death.

First incarnation of preprints should have the same requirements as current

journals. More could be added later.

What about data fabrication screens? Journals have an ethics system using

screening tools or reviewers are picking it up. Who is going to screen this

for preprint servers?

> We actually ended up having a longer discussion on how a preprint

> server might be paid for. BioRxiv claimed that this is not a problem.

> Based on my arXiv experience, I am a little more skeptical.

What is the economy of this? Non-commercial central repository? Problem with

any private ownership is that it could be sold.

One archive or not? BioRxiv could scale up and thinks can get funding to deal

with a much larger demand.

Could preprint server system offset tendency in China to move to Chinese

language journals?

It would be harder for Chinese scientists (cultural barriers) to participate

(through submissions) in a preprint server system.

Is there anything that can be done to make it even easier for China/India/etc

to submit?

Federated funding model is most likely for preprint server.

Databases are counter example - typically funded by single funding agency.

> Another question to be considered:

What tools can help readers to find material (topic, date, investigator,

quality, viewership, significance, novelty)?

"Like" button?

Bibliometrics is somewhat mature fields.

Download numbers do not correlate with paper quality.

> And another idea not so centrally related to the purview of the group

> in the final minutes:

Private sector could come up with solutions for organizing preprints.

Preprints are good for economy since they raise access to literature for

biotech companies.