Building a Digital Ecosystem Breakout Session
BD2K AHM Friday, November 13th 2015

- What are the questions and points of discussion regarding The Commons, and how this will play a role moving forward?
  - What would be early successes for the Commons?
  - What are we missing in our concept of the Commons?
- What are the shared resources each Center or other BD2K-funded initiative will produce?
  - What software and/or data that could be shared in the Commons?
  - What features in the Commons would facilitate the sharing of resources?
- What resources beyond NIH funded research should be considered as part of the digital ecosystem?
- Are there good models we can learn from, either in biomedical sciences or other disciplines? (Nationally and internationally)
- Who are the contributors, users, and ultimately, the end stakeholders for this effort?
- Your recommendations of additional elements.
- How do digital objects get prioritized and validated to be included in the digital ecosystem? Should we not worry about junk in/junk out because there will be enough data to eliminate noise? (Google model). May have challenges with genomic type of specialized datasets.
- What are the minimal set of metadata for each digital object in Commons? #CEDAR #biosharing

Slide Deck generated from AHM and updated by Chairs. These slides capture the essence of the discussions and address the charge questions above. We did not capture the names of attendees. The slide text is captured below:

Is "ecosystem" the right analogy?

- Ecosystem ➔ Darwinian view: survival of the fittest
- We want a **civilization** not a 'dog-eat-dog' ecosystem. Example ideas below:
  - "**Town planning**" as a better analogy? Create framework (streets, utilities) within which collaboration and competition can occur
  - GPS as a navigation system
  - **Gardening:** the **farmer** who creates the conditions for many flowers to bloom

How might we ensure success for the Commons?

- **Engage training programs**: ensure that data/software/tools used in training are available in the commons; support **trainees** to engage with all elements of the commons ("eat own dogfood")
- Provide a **roadmap** that people can use in planning
- We need '**champions**' that use all parts of the digital ecosystem and can be both advocates and trainers for others.
- **Vignettes** for different workflows, pipelines for analysis, to answer different biomedical/biological questions

What are we missing in our concept of the Commons?

- Human **expertise** as a vital resource
- **Curation** / quality control of data so that quality improves over time
- The commons is not just a place where data and tools sit, but a locus for **transactions**
- **Fractal nature**: not just national, also local/institutional
  - Everyone should be building (components of) the Commons, and at different scales
- Clear statement of target stakeholders
  - One person's data is another person's results. A digital ecosystem allows making data sharing and resuse possible.

How can we measure success?

- Create a roadmap the phases the development and expectations for The Commons
- Measure all **transactions** involving BD2K components: commons, training, centers, software, standards, e.g.:
  - User A uses data from B and C
  - Trainee E uses code from E
- New results that link disciplines
- **Reputation:** community-based evaluation scheme that is more organic than simply usage/citation statistics
  - Ask The Commons user, were you successful in doing what you wanted when you used it? And what can we do to make it easier

Are there good models that we can learn from?

- IBM Watson: provides easy access to test datasets and Watson software
- Google Environment and Apple that have human factors that evaluate and demonstrate interoperability (eg between different Apps)
- Amazon, Netflix or eBay ratings and recommendations for products and people
  - Create a memory of activities on digital objects
- Other systems are being built in other fields (e.g., SEED, MG-RAST, iPlant, Earth System Grid, kBase)—we should talk to them

Background material: see https://datascience.nih.gov/commons

Breakout session notes

- Can we select potential prototypes that will help with the digital objects that will go into The Commons.

- Data and the tools are “inseparable.” Once Vivien has the datasets (e.g., microbiome), then we have other groups add analytics/tools that work on the data.
- The prototype should demonstrate the use of The Commons by the broader data science community. Ensuring “productive” access so it’s not just a matter of data access, but supporting the transformation of this into knowledge.
- Are there other things that should be aggregated/associated with the data that is going into The Commons?
- Who is The Commons trying to be useful to? It will probably be different types of users (e.g., different use case actors). → large range (clinicians, scientists of different disciplines, etc.). and ultimately to serve the NIH researcher → perhaps identify user groups early on who we can look to for input on this?
- Can we embed trainees from the Centers, perhaps also through the TCC, that will be trained in The Commons as dedicated tool users, working with specific BD2K datasets, DDI, and the Cloud credit model, etc. → These individuals will be extremely valuable and know how everything works. → Talk with Jack Van Horn. Activities in this platform would be helpful and coordinated via the TCC.
- We need “champions” that use all the different parts of the digital ecosystem and then can be advocates. It could be postdocs or other trainees. Start to stand up right now some individuals who put all the pieces together and do this full-time → and then have them go back to the training programs as a data scientist.
- The challenge will be The Commons is evolving; is there also a role such individuals can play in also helping to define The Commons to shape it as it moves forward.
- It would be useful to have vignettes for different workflows, pipelines for analysis, to answer different biomedical/biological questions. These are extremely useful and powerful for people to understand The Commons and all of its anticipated parts → is there a way for people to contribute these vignettes, and perhaps be used towards publications?
- A layer that enables people to connect together via datasets, tools (discovery and sharing).
- What are the “survival” and ‘fitness” criteria for the ecosystem (comparison to a natural, biological environment).
- What are all the other parts and resources of the ecosystem, as we put it together?
- One person’s data is another person’s results. A digital ecosystem allows making data sharing and resuse possible. The problem is that we often don’t have an ability to verify the data that we have in shared repositories (there is no motivation to fix data in the database as time goes on → sustainability and curation issues).
- How do we promote synchronization imaging and genomic data, for instance? Can The Commons help with these types of issues (cross-modality; and human vs. animal modeling). Can it help with coordination?
- Right now, we don’t have a digital ecosystem → we are the ones who are going to create this new environment (via The Commons), and so we need to be conscience of what we are putting into The Commons and what shape it will take.
- Metasystem that shows data that isn’t connected and to see what is going on → tools that for some reason aren’t a part of the overall ecosystem, for instance.
- Example digital ecosystem might be IBM Watson; look at how it was setup and how they have made it usable/learnable. Another example would be eBay, which has meta-levels of information (community-based trust/reward/reputation systems).
- What other things can we learn from so we aren’t doing this de novo.
- Reputation: community-based evaluation scheme that is more organic than simply usage/citation?

- Are we okay with an ecosystem where “dog eats dog” or do we want a civilization that is more planned? Perhaps set an initial plan with rules, and then let it evolve. All the users must be contributors from the start. Make the users a part of the ecosystem itself, and this could also make it a more democratized framework for data access/analysis.
- The interoperability between different silos that have been developed; The Commons should enable the interoperability of different Centers, tools.
- Training on “common” datasets → real datasets need to be built so that the training programs can take advantage of them.
- The more global digital ecosystem really also needs to look at the “micro” and “mini” levels (departments, institutions) for sharing. We often let people get away with not following up on curation, annotation, etc. that is in the data sharing statements. Can we make this a part of Offices of Research; data management plans should address ultimately the Commons -- but do it locally first, too. → NSF has a strong requirement and setup in this regard (but there is no evaluation of follow-up)
- If you have a data sharing plan, make sure you put budget into it.
- Can we create a roadmap the phases the development and expectations for The Commons, data scientists, etc. usage? We also do not want to build a bridge to nowhere, we need to know or have some idea of what we do want to achieve. Example questions that The Commons will handle. Pose a set of problems (like the AI paradigm)?
- The Commons can enable more advanced, sophisticated methods for data science; the low-hanging fruit of analysis will already be done by the time data gets into The Commons.
- Analogy to creating a GPS system.
- A lot of people would like to share their data, but they simply don’t know how to.
- “Bring your own data” (BYOD) as part of the exercise for trainees in the use of The Commons.
- The Commons is meant to be a sandbox for exploring data and working with it; it is not meant to be a data warehouse and long-term archive. ← This is probably true for the cloud components; but are there more stable resources (e.g., for long-term storage?).
- The ecosystem should grow, but certainly not the number of web sites/access points.
- Define some activities that we do regularly, and that The Commons will make better or (even enable/make possible)
- Like how Amazon Videos or Netflix make suggestions: these are some similar tools or analyses (or subsequent actions) that others have done or also used. A “memory” of activities on given digital objects.
- Measures of success: transactions between the different components that make up The Commons (using DDI and then applying it for training or in the Cloud Credit pilot).
- The Science of Team Science is grappling with similar evaluation metrics; things like the number of papers, etc. are conventional and do not work necessarily. What if we could look at the scientific discoveries that are made as a result of the effort?
- Other efforts like the Google environment, or Apple, must have some type of human factors evaluation that demonstrates the degree of interoperability (e.g., between different apps). But we don’t have a similar set of metrics for (team) science.
- Ask The Commons user, were you successful in doing what you wanted when you used it? And what can we do to make it easier?