Workshop report 'Solving Bottlenecks in data sharing in the Life sciences'

Open PHACTS / GEN2PHEN workshop September 19 and 20, 2011

Aims:

The workshop aimed to explore two directly related topics: data sharing (pushing data outwards from its source to one or other online resources) and data access (pulling data from one or other online resources for additional use). The goal is to maximally enable and promote 'open' data sharing, i.e., precompetitive, unencumbered, unrestricted, universally equitable dissemination of datasets generated by academia and (to a certain degree) by industry. Therefore, focus was restricted to datasets that are ethico-legally 'safe' to share, or can be made safe to share by some pre-processing, aggregation, anonymisation or via advanced data access methods designed to protect the data. It is accepted that certain other datasets will not be possible to share in any kind of 'open' fashion.

Presentations on Day 1

Three short presentations by data owners about 'their issues': Evan Bolton from PubChem, Frank Schacherer from Biobase, and Christine Chichester from Open PHACTS.

Keynotes:

´Introduction of IMI Ju`, Ann Martin, Principal Scientific Manager Knowledge Management

'Open Access and Open Source: no free lunch!' Jan Velterop, Open PHACTS and AQknowledge

'Open Source licensing and sustainability models for effective data sharing in the Life Sciences', John WIlbanks, Creative Commons and SageBioNetworks

"Forms of OPEN Sharing that avoid data disclosure, and methods to make CONTROLLED sharing equivalent to OPEN sharing". Anthony J Brookes, GEN2PHEN Co-ordinator

Day 2:

Workshops on the legal, sustainability and social challenges to be addressed to build a culture of data sharing

Outcomes from the Workshops

Summary

The meeting was well attended, signalling considerable interest in the topic. This probably also indicates that a kind of a tipping point is being reached regarding the sharing of data as a common and expected activity. This itself is clearly driven by the emergence of -omics technologies and large scale studies, which demand new research business models and data management structures. These things, however, are typically not the expertise (nor the interest) of scientists, but require specific professional knowledge.

A related matter that is tightly intertwined with any consideration of data sharing and data access is that of sustainability of online data resources. In short, how can resources be made sustainable if the primary commodity (the data) are deemed 'valuable' and so often not shared to the online resource in the first place and yet have to be handed on openly (here, meaning 'free') by that resource. These issues were debated in the workshop, but those discussions did not extend to the question of sustaining the data generation activities themselves, nor to enabling the radical further development of online resources. It was felt that sustainability challenges cannot be resolved in a ‘silo-ish’ manner, and so a coordinated approach to the problem is desired, with exemplar sustainability models being part of the way forward. As part of this, funders need to allocate more money to data stewardship in research grants, but only distribute those funds given evidence of actual delivery on promises (via local or outsourced approaches). Lastly, projects themselves need to realize that starting sustainability efforts cannot be left as a low priority or started too late to be in effect once the lifespan of the project is over.

The results of the workshop will be addressed in a general perspective summary paper or publication, targeting Human Mutation

The general conclusions will also be used by the participating sponsoring projects to inform future project approaches.

Summary OPS / GEN2PHEN Workshop discussions and suggestions

(More detailed info in the addenda)

A. Introduction:

The workshop dealt with three closely inter-related and inter-dependent topics:

1. enabling data sharing (getting data 'out' or 'exposed' from source)
2. providing data access (reaching in to get the shared data)
3. creating sustainability (challenges and opportunities that stem from the innate 'value' of data, and the maintenance and development of data sharing platforms)

B. Main problems identified

B.1. Legal/licensing

a) No actual jurisprudence on new forms of data publication.
b) Operating open sharing in the complicated legal landscape is extremely challenging.
c) How are datasets supplementary to publication covered by copyright?
d) Combination of data sets with different licensing conditions is a challenge.
e) Access rights to back- and foreground often after the project; new data from old data.
f) ETC.

B.2. Social/Ego system:

a) No incentive model for sharing.
b) No citation possibilities.
c) Unclear trust level of databases.
d) The prisoner's dilemma of data sharing: advantageous for the group if everyone shares. Advantageous for members of the group to not participate in sharing and reap the benefits of others sharing.
e) How do we protect against database "Piracy" - trust that the data will not be grabbed without due recognition, attribution or citation in publication as paper or web application.
f) How will data sharing be monitored for proper attribution and citability crediting.
g) ETC.

B.3. Sustainability

a) To prevent waste (duplication of effort, data loss) and to optimize knowledge discovery (linking data), existing large-scale data resources and innovative software applications and services need to be accurately evaluated, prioritized, and traded in a speedy, fair and trusted environment.
b) Presently, these transactions are frustrated by (1) a lack of interoperability (data standards), and (2) a burdensome legacy of incompatible patch-work licensing (slowing rather than fostering innovation), (3) lack of enabling infrastructure to make this an economic process
c) Solving these problems from the top-down is likely to fail as this involves negotiations between a large number of independent stakeholders hoping to resolve conflicts between dozens of licenses and data formats.
d) Increasing sustainability costs are challenging existing funding priorities which prioritise shorter-term research projects over long-term maintenance funding (which itself may cause problems, squeezing out innovative approaches)
e) Just developing a sustainability plan for Open PHACTS (and GEN2PHEN and other related projects) in a vacuum may be insufficient.
f) A central authority will never be able to broker all N X N licensing possibilities.
g) ETC.

C. Recommendations

C.1. Legal/licensing

a) Keep things simple: use a single, standard, simple, globally accepted license for this.
b) Create machine-.readable resolution service for data licenses.
c) Create stack documenting the current procedures and data interoperability guidelines.
d) Create market place where data owners and comsumers can use one click solutions for agreements/licenses from the stack.
e) Use "MetaPHACTS" – a hard core of open, linked, rdf-ed data which can be re-released under a single, simple licence (e.g. CC-BY, …) and all other sources have the option to include a link ("meta-phact") that indicate the existence of other data in the right place (obviously the "meta-phact" must also be under the single, simple license opted for).
f) There's a second choice here – the 'hairball' of unlicensed data. Keeping all data in, but saying the license is different/unknown is an approach which is currently working reasonably well for both ChemSpider and PubChem.
g) Develop a regime for managing liability for data providers, and data sharing infrastructure
h) Develop strong norms that new knowledge creation is the entire point of OPS/related/similar projects.

C.2. Social/ego system

a) Enable the data ecosystem of the future.
b) We get from the current state to the desired future state by removing barriers to sharing data by promoting advantages to data owners'.
c) The ecosystem of data producers is served by recognizing sharing. Sharing can be called "publication". Establish similar processes to paper publication -- citation mechanism, "databank" -- a measure of the utility of the data to others -- not an impact factor, but more like Google page rank.
d) All secondary database usage must declare provenance of the primary data being disclosed.
e) Citation, impact factor (need of impact factor for data citation). Datasets are part of the scientific output
f) Appropriate use of BY attribution to help the above (while recognising John Wilbank's comments that attribution does not equal citation)

C.3. Sustainability

a) The sustainability plan must be developed AND implemented in a broader context (i.e. all projects need the same so interoperability gets another meaning in this context: avoid re-inventing of wheels or developing multiple legal and sustainability solutions. But we also heard from Biobase that one size may not fit all in terms of viable business models. Different domains (as now) sustain different models.
b) The sustainability plan must be developed in sufficient detail and made open to the community to invite support and engagement. Alternatively, we were encouraged to try a multiplicity of different approaches which may not result in a single clear approach early on
c) Creative business models must be developed to generate revenue streams - open is not necessarily free - depends on costs and value add for content and service.
d) Open sharing system must gain support and endorsement from community organisations such as Pistoia Alliance, Concept Web Alliance, VIVO, ORCID and other related (IMI) projects dealing with datasharing and sustainability.
e) Develop solutions as an online market place (for example: The Data and Knowledge Exchange (D&KE) as a virtual online market place that presents information technology 'goods' and 'services' for trading (e.g., data, software, workflows, etc).
f) Such market place should provide a transparent platform and standards for producers and consumers of knowledge resources to meet and negotiate the terms of data/software exchange.
g) This market place approach should drive data sharing (openness) and convergence on data standards (interoperability).
h) Involve industry beyond pharma.
i) Impact and community engagement: Develop Apps that generate alerts on topics of our interest.

j) Suggestion to include a section in project proposals on what consortia are going to do with the data and line item for data stewardship and the associated cost.

D. Conclusions and suggestions

a) Common denominator in the keynotes and discussions in the breakaway groups was that kind of a tipping point has been reached and that the 'vacuum' drawn by changing the way science is done (i.e. knowledge is published/shared) needs to be filled with new business models and data sharing agreements.
b) The extensive number of challenges stated and solutions recommended need extensive studying to be integrated into one proposed approach which suggests the need for a dedicated team and not a 'Friday afternoon' approach by already overcommitted project team members/coordinators.
c) Designing and implementing such a broad sustainability plan is typically not the expertise (nor the desire) of scientists and requires specific professional legal and business model knowledge.
d) In addition, most data intensive projects have the same sustainability challenges, which should not be resolved in a 'silo-ish' manner, so pooling resources and assigning this task to a professional team to be resolved and implemented for all projects would make sense.
e) Projects need to realize that starting sustainability efforts cannot be made a low priority or started too late to be in effect once the lifespan of the project is over.
f) Although it is recognized that each project has its own goals and deliverables, Open PHACTS and GEN2PHEN leadership, in collaboration with related (IMI) projects could take a leadership role in organizing such a dedicated team of professionals to get the job done.

Addenda -

Notes from the individual workshop teams

Results of the workshop discussions on boat 1

Challenge 1: How do we get sustained submission of data to an open sharing system?
→ Solution to challenge 1:
1. Attractive incentives recognition and reward for contributions
2. Easy submission process
3. Never have need to re-submit
4. Longevity of data guaranteed

Challenge 2: How do we motivate a scientist to submit data to the open sharing system?
→ Solution to challenge 2:
1. Utilise ORCID ID to identify registered submitters
2. Show submitter contributions in a dashboard view
3. Reward contributions by submitter:
    - Dashboard link for submitter's CV
    - Contribution points earned to gain benefits
    - Dashboard view of all submitters to generate community

Challenge 3: how do we provide incentives to share and value knowledge/ data/ tools?
→ Solution to challenge 3:
Create an Ebay-like trading place. It should contain a Semantics based search engine, and flexible licensing/ attribution conditions.

Challenge 4: how do we create sustainability for individuals and consortia?

→ Solution to challenge 4:
Create a Ebay-like trading place. Success/ ranking on the trading place provides sustainability. OPS sustainability is likely to be as an integrated data service provider and/or a governing body and/or the provider of the semantic search engine for the matchmaking.

Challenge 5: How can we require data publication with context and provenance of both academic and corporate data producers?
→ Solution to challenge 5:
Require publication of data in contract terms as a condition of payment. Require publication of data for publication of papers. Require publication of data for approval of new drugs.

Challenge 6: How can we identify and fill gaps in data holdings available for reuse?
→ Solution to challenge 6:
Using the dashboard, perform a gap analysis to identify missing data (examples include textbook knowledge, phenotypic data for normal subjects across the world); agencies then fund projects to fill the gaps.

<u>Results of the workshop discussions on boat 2</u>

Social challenges: why people do not share / what would be the incentives for sharing
→ Solution to social challenges:

- Benefits (carrots):
    1. It avoids duplication
    2. Citation, impact factor (need of impact factor for data citation). Datasets are part of the scientific output.
    3. Potential data curation
    4. Reasonable data format
- Potential solutions (sticks)
    1. Project funders require that data is published. *Create an advocacy group for promoting data publication.*
    2. Publishers require that supporting data is made public and available for a long time.
    3. Suggestion to include a section in project proposals on what consortia are going to do with the data.
    4. "Black list"
    5. Create trusted data stewardship centres
    6. Develop a quality seal on shared datasets
    7. Re-use of data should be rewarded
    8. Willingness to collaborate with NCBI

Sustainability challenges
→ Solution to sustainable challenges:

1. Involve industry other than pharma (without losing focus) – side, parallel lines
2. Really go open source (problem of the background IP). Create the environment for an open source community.
3. Impact and community engagement. Develop Apps that generate alerts on topics of our interest.
4. Build something small that works and then go to the big guys.
5. Make our goals open and public.

<u>Results of the workshop discussions on boat 3</u>

Challenge 1: Combination of data sets with different licensing conditions

→ Solution to challenge 1:

1. Go back to the people that own the background and ask them for a new license for the Open PHACTS Consortium – in return they get an expanding database and can use the platform to point to their data
    - Use a single, standard, simple, globally accepted license for this
    - Create machine readable resolution service for data licenses
1. Create MetaPHACTS layer for non open or unlicensed data
    - Two corpuses: unambiguous open + complex-licensed
    - For new data, unambiguous policy to go into open corpus

Challenge 2: access rights to back- and foreground after the project; new data from old data

→ Solution to challenge 2:

1. Develop a regime for managing liability for data providers
    - access control: due diligence, liability, government lobbying
    - Insurance: business opportunity
    - Policy: assumes provenance (levels of attribution), already clear it in the project proposal at the beginning of the project
2. Develop strong norms that new knowledge creation is the entire point of OPS
    - Reward people who ask good questions of old data
    - Provide credit where credit is due for data providers

Challenge 3: Who are the stakeholders in a sustainability model? Who pays for what?

→ Solution to challenge 3:

1. Create a market place for open innovation, stakeholders are public funding parties, foundations, pharma companies, companies building on the data, patient groups, academic groups
2. Run different business models in parallel – create a market place for all 3 levels (digital data preservation, data formats and standards, knowledge level)

Challenge 4: How to attract data owners / data sets?

→ Solution to challenge 4:

Giving them recognition for their data, the possibility to match and integrate and to generate more traffic and usage of their data set

Challenge 5: How to create incentives for individuals, institutions and funding agencies?

→ Solution to challenge 5:

1. Citation possibilities for data; link to individual scientist, e.g via ORCID
2. More collaboration / networking possibilities
3. Publish now and release later
4. Placing data sets in OPS implies a certain quality stamp of approval
5. Join early – later it becomes more difficult and/or more expensive

Challenge 6: Once we have the incentives, how do we reduce the workloads?

→ Solution to challenge 6:

1. Come up with technological solutions that make submission a one click effort
2. Expert help in institutions out of grant overheads

Photos by Rob Hooft (http://rwwh.smugmug.com/Company-Gatherings/Open-PHACTS-and-Gen2Phen-2011)