The future of annotation/biocuration

Amos Bairoch

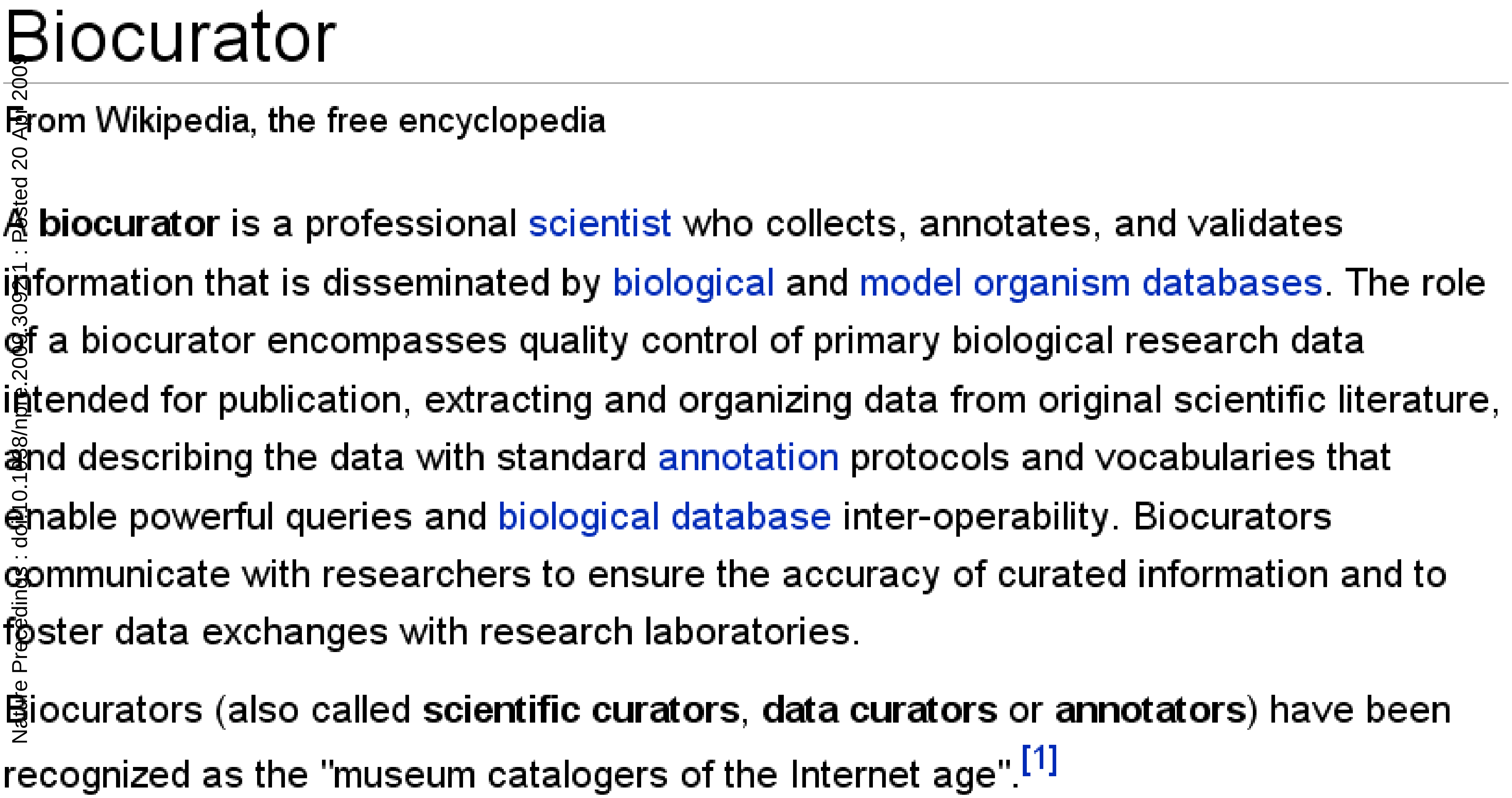
Swiss Institute of Bioinformatics (SIB) and University of Geneva

This is not a talk!

- But rather a series of truisms (obvious facts);
- Most of which have already been said by Janet yesterday evening and a few minutes ago by Phil;
- But maybe some are worth repeating!

First problem: defining the scope of annotation/biocuration

- There is a huge range of activities that are put into the framework of annotation;
- There are various levels (depths) of annotation: going from assigning a gene name to a newly sequenced genome to the creation of a whole resource around one single biological entity;
- The main guiding criteria is that all of those activities are meant to help the Life Sciences community to make sense of all the data that is accumulating.



Biocurator

From Wikipedia, the free encyclopedia

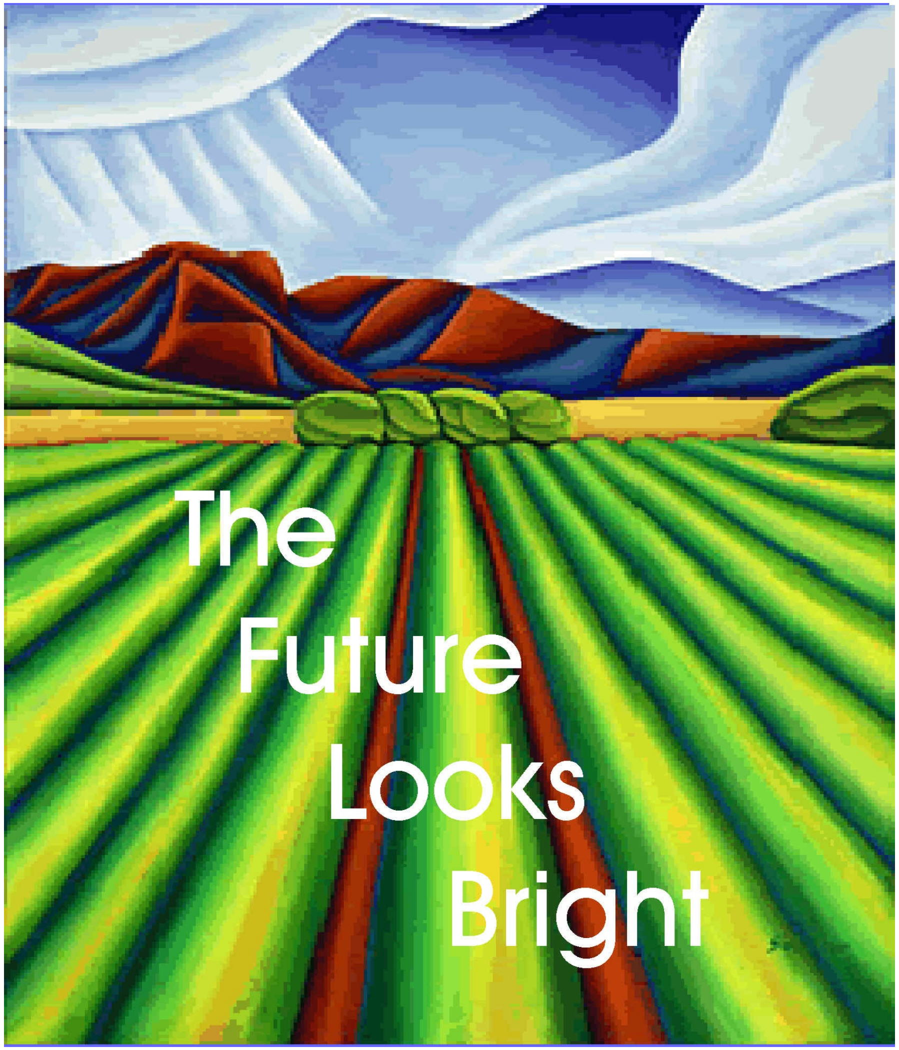
A **biocurator** is a professional scientist who collects, annotates, and validates information that is disseminated by biological and model organism databases. The role of a biocurator encompasses quality control of primary biological research data intended for publication, extracting and organizing data from original scientific literature, and describing the data with standard annotation protocols and vocabularies that enable powerful queries and biological database inter-operability. Biocurators communicate with researchers to ensure the accuracy of curated information and to foster data exchanges with research laboratories.

Biocurators (also called **scientific curators**, **data curators** or **annotators**) have been recognized as the "museum catalogers of the Internet age".[1]

Museum catalogers?

- I believe this is a misleading analogy: biocurators are much more **curators** than **catalogers**;
- Museum curators were/are researchers that have made a tremendous contributions to the natural sciences;
- As long as we are seen as catalogers, the value of what we provide is not going to be judged by its intrinsic value.

So, what about the future?



From flow to flood to…

- We used to speak an increasing flow of data, then it became a flood, recently a tsunami;
- We need to come up with a new word for what is awaiting us!;
- This is not only true for sequencing data, but in an increasing number of other technology-driven area such as proteomics, microarrays, imaging, etc.;
- Increasingly, this data will not have been generated to answer a specific biological research question but as collateral (damage?) to routine diagnostic, quality-control and environmental-monitoring procedure;
- Consequently no budget will have been allocated to curate this data.

So we all know that…

- Nobody will ever be able to manually annotate all the macromolecular biological entities that exist on this planet;
- And consequently that automatization is the only solution;
- But you can not propagate something that does not exist;
- Therefore corpora of high-quality manually-annotated data are an essential requirement of the worldwide biocuration efforts.

Manual annotation is costly

- Example: *Schizosaccharomyces pombe*;
- We just completed in UniProtKB/Swiss-Prot the annotation of all 4’957 proteins (5’001 genes);
- I estimate it took 10 women/men years to achieve this goal;
- That translate into 500 entries per year per FTE;
- This was done by:
  - Reading about 1’600 pombe-specific papers;
  - Carefully propagating information from many other eukaryotic organisms (yeast, mammals, Drosophila);
  - Using a variety of external resources (GeneDB_Spombe, SGD, PDB, etc).

And another similar observation

- Swiss-Prot is soon going to reach the 500K entry level;
- About 50% of these entries have been manually annotated and re-annotated and re-annotated and…..
- So what about the second half:
  - They have been provided by the HAMAP pipeline;
  - Probably one of the most reliable (“ better safe than sorry”) automatic annotation pipeline operationally deployed;
  - Which so far is only used in the context of bacterial and archaeal entries but which is progressively going to be used for eukaryotes.
- The 250K manual entry required a cumulative to total of 600 w+m/years (over a 23 year period);
- That computes to 420 entries per year per FTE.

But this not all

- The previous FTE figures represent only the direct cost of biocuration;
- But biocuration can not be done in a vacuum it requires:
  - The development of software platforms to allow the process to be efficient and standardized;
  - The development of interfaces (mostly web-based) for users to make us of the annotation;
  - The development of controled-vocabularies and ontologies;
  - IT and administrative support.

Community annotation

- Up to now, it has not been a big success!
  - Wikis: easy to deploy, not used as much as one would have thought they would have been, given the success of Wikipedia;
  - Adopting a protein: many potential parents, but when it come to take care of the children…
  - Miscellaneous other schemes: very limited success.

Why?

- Because there are currently no direct incentive to researchers to spend some of their precious time helping in the curation effort;
- In fact there are generally much more pitfalls then incentives;
- And we, as a community, are also partly to blame:
  - We do not spend enough time teaching young scientists how our resources should be used and how by contributing they will help their own research;
  - There are a probably more than 2’ 000 resources available: how can we expect them to know what they should submit to whom?
  - Until recently we have not done enough to work in close collaboration with journal editors and publishers (conversely: they have also ignored us for a long time!).

Rewards and punishments

- We need rewards and punishments to make researchers active actors in the biocuration process:
  - If journals agree:
    - Fast track publication;
    - Publicizing how many times a given paper is accessed through a link originating from a database;
    - Refuse to publish if the relevant data has not been submitted to a database (this used to work for DNA sequences, but is less enforced than it used to be).
  - If funding bodies agree:
    - To oblige funding requests that include significant data generation to include funding for the curation and storage of this data.

Semantic tagging

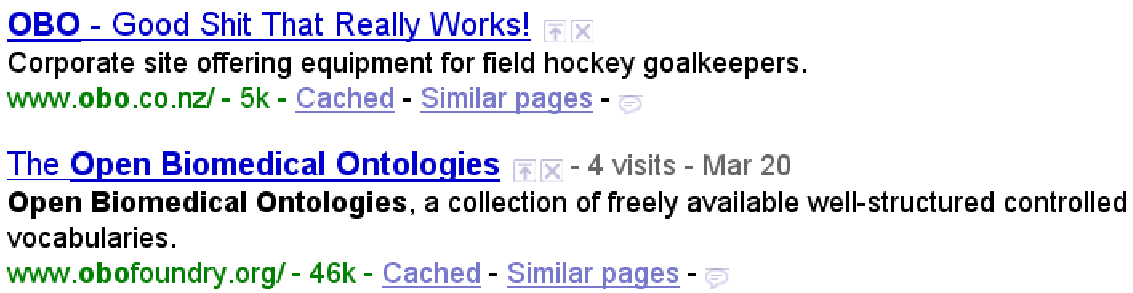
- We are all eagerly waiting for the time when articles will be submitted ready to be “ automatically” processed to extract the relevant facts;
- It is quite depressive to think that we are spending millions in grants for people to perform experiments, produce new knowledge, hide this knowledge in a often badly written text and then spend some more millions trying to second guess what the authors really did and found.
- Yet, lets not hide that the biggest problem is that often these are not facts but inferences and often also wishful thinking;
- Semantic tagging will be a significant step forward but….

Lets be honest…

- Biocurators are often more critical and have a broader view than the average lab scientist;
- We often spend more time “ de-annotating” what people have reported then entering their data;
- We all know that, but we are shy of making this known outside of our field!
- Journals should make more use of the collective knowledge of annotators before accepting a paper for publication.

Standards

- Yes obviously we need ontologies. Everyone agrees and the good news is that they are thriving. Just look at the list on OBO and you will see how active this field is!



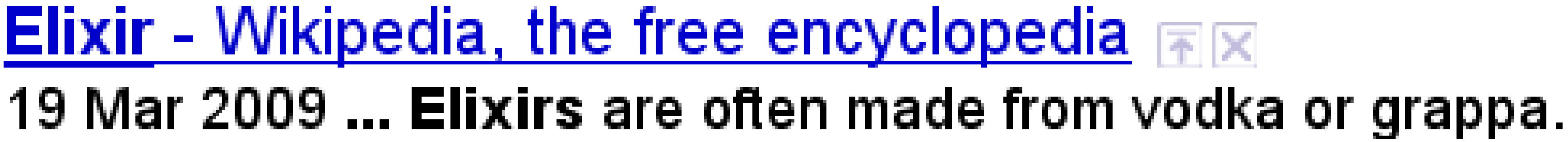
OBO - Good Shit That Really Works!
Corporate site offering equipment for field hockey goalkeepers.
www.**obo**.co.nz/ - 5k - Cached - Similar pages -

The **Open Biomedical Ontologies** - 4 visits - Mar 20
Open Biomedical Ontologies, a collection of freely available well-structured controlled vocabularies.
www.**obo**foundry.org/ - 46k - Cached - Similar pages -

- But lets not forget that we also need less prestigious and therefore often less prone to be funded efforts to develop:
  - Controlled vocabularies
  - Nomenclature
- For example: the enzyme nomenclature committee has, since the 60s done a tremendous work in putting order into what is known as the EC system, yet this effort have never been directly funded.

Funding…..

- We are a pain in the neck for funding bodies: we require long-term solutions and not 2 to 5 years research grants;
- They recognize our efforts but rarely have the tools to allow infrastructure to be funded;
- This is why efforts such as ELIXIR are important to the future of our collective efforts (and elixirs are good stuff!)



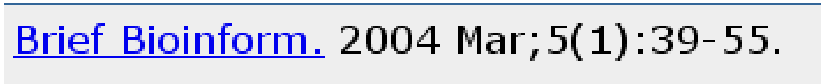
Elixir - Wikipedia, the free encyclopedia

19 Mar 2009 ... **Elixirs** are often made from vodka or grappa.

What to learn from physicists

- Physicists have been successful in securing funding for large scale research infrastructures (CERN, synchrotrons, etc);
- They are entrenched in all levels of decision processes in allocation of research budgets;
- We can’ t blame them for being efficient!;
- They have learnt a lesson we have not yet completely assimilated: union makes strength;
- We need to act together and not compete for scarce financial resources;
- It’ s a difficult exercise: one need to strengthen large centralized resources (EBI, NCBI, etc) without killing in the egg worthwhile existing or emerging efforts spread over a wide variety of institutions worldwide.

Six observations to databasers



Brief Bioinform. 2004 Mar;5(1):39-55.

1. Your task will be much more complex and far bigger that you ever thought it could be.
2. If your database is successful and useful to the user community, then you will have to dedicate all your efforts to develop it for a much longer period of time than you would have thought possible.
3. You will always wonder why life scientists abhor complying with nomenclature guidelines or standardisation efforts that would simplify your and their life.
4. You will have to continually fight to obtain a minimal amount of funding.
5. As with any service efforts, you will be told far more what you do wrong rather than what you do right.
6. But when you will see how useful your efforts are to your users, all the above drawbacks will lose their importance!

And the future is also…

- The International Society for Biocuration;
  - To lobby for much more resources being put into biocuration;
  - To foster exchanges between biocurators;
  - To help organizing conferences such as this one;
  - To contribute to the awareness of recruiting bodies that biocurators are not failed researchers playing around with computers, but rather experts with a broad knowledge of the issues and intricacies in the Life Sciences.

So: do not forget to join the society!!

ISB International Society for Biocuration