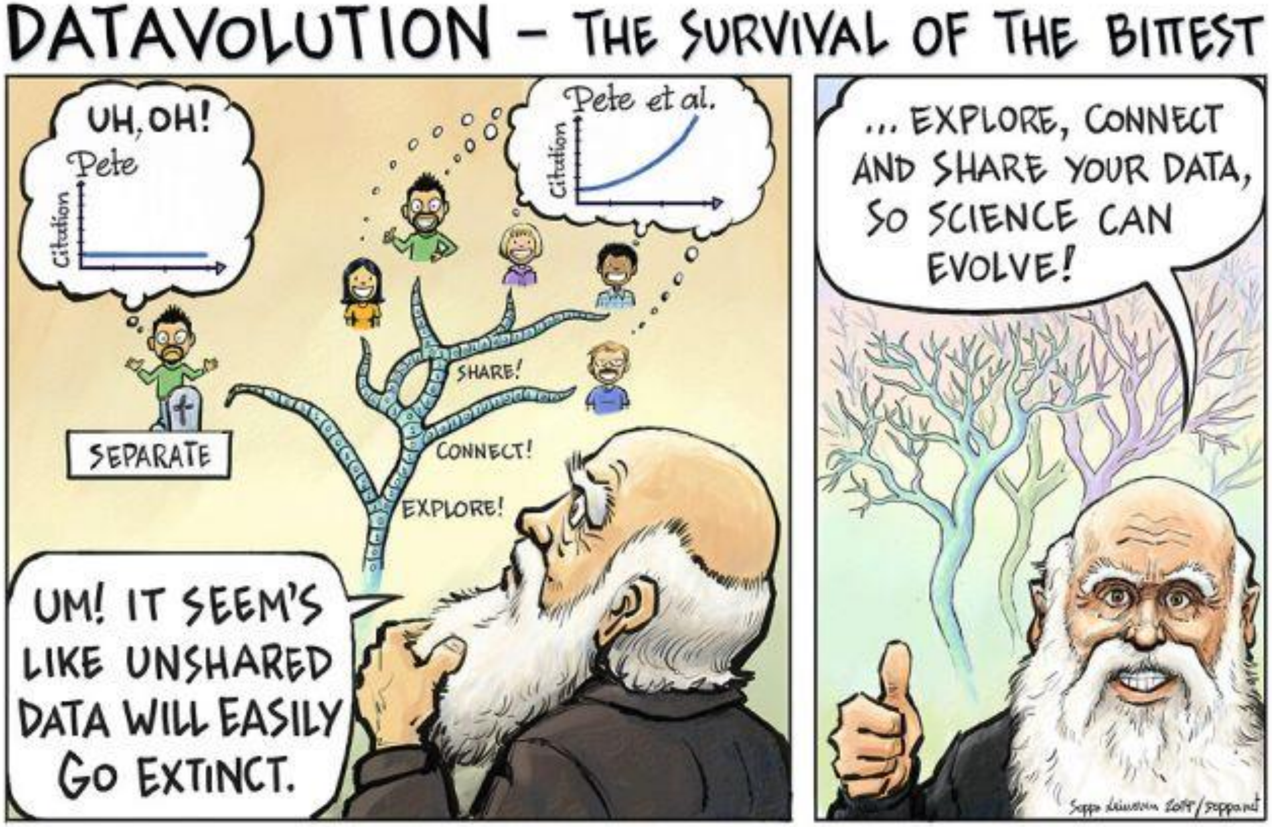
Introduction to Data Resources for the Life Sciences

Helen Berman

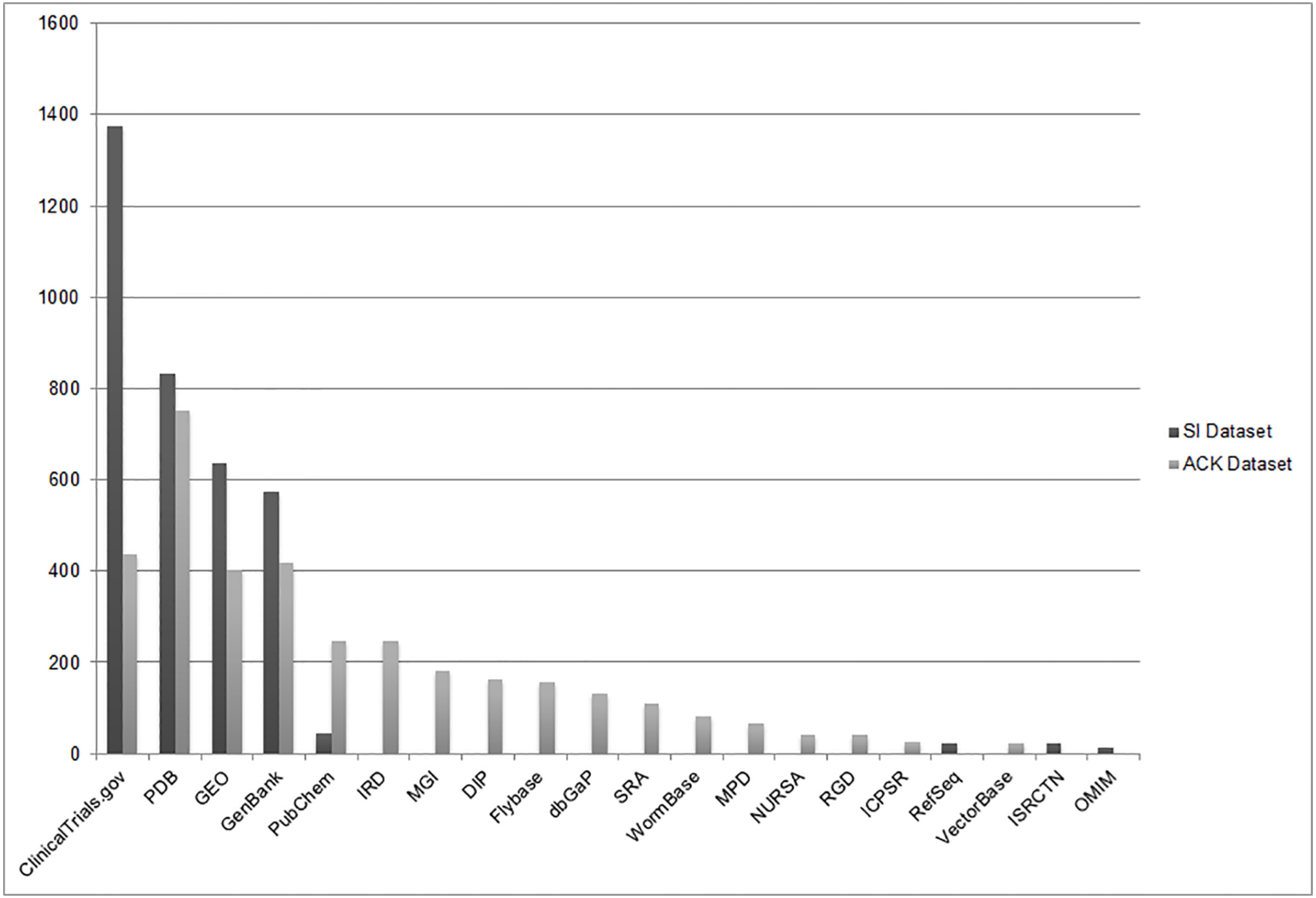
HFSPO Strasbourg
November 18-19, 2016


DATAVOLUTION - THE SURVIVAL OF THE BITTEST
UH, OH!
Pete
Citation
SEPARATE
Pete et al.
Citation
SHARE!
CONNECT!
EXPLORE!
UM! IT SEEM'S LIKE UNSHARED DATA WILL EASILY GO EXTINCT.
... EXPLORE, CONNECT AND SHARE YOUR DATA, SO SCIENCE CAN EVOLVE!
Seppo Leinonen 2019 / seppo.net

Future Support of Data Resources in the Life Sciences: Proposed White Paper

- Which life science data resources should be included?
- Which parameters define the boundary?
- What kind of support/funding mechanisms are needed?

Fig 4. Repositories identified from the PubMed SI field and PMC Acknowledgements where datasets were deposited.



Read KB, Sheehan JR, Huerta MF, Knecht LS, Mork JG, et al. (2015) Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. PLoS ONE 10(7): e0132735. doi:10.1371/journal.pone.0132735
http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0132735

PLOS | ONE

Types of Data Resources

- **Archival data resources**: primary data on which other data resources are built
- **Specialty resources:** expert curation in a focused area
- **Knowledgebases**: integrated resources containing annotation from many different resources
- **Value-added resources**: extensive computational annotation

Issues to Address

- Funding model
- User community
- How are data used and how do you measure?
- What is the impact and how do you measure?
- What would be the impact if the resource no longer existed?
- What is the contingency plan if support is lost?
- What are challenges to financial sustainability

Funding Model

- Types
  - Government
  - Foundation
  - Submission fees
  - User fees
  - Membership
- Gaps

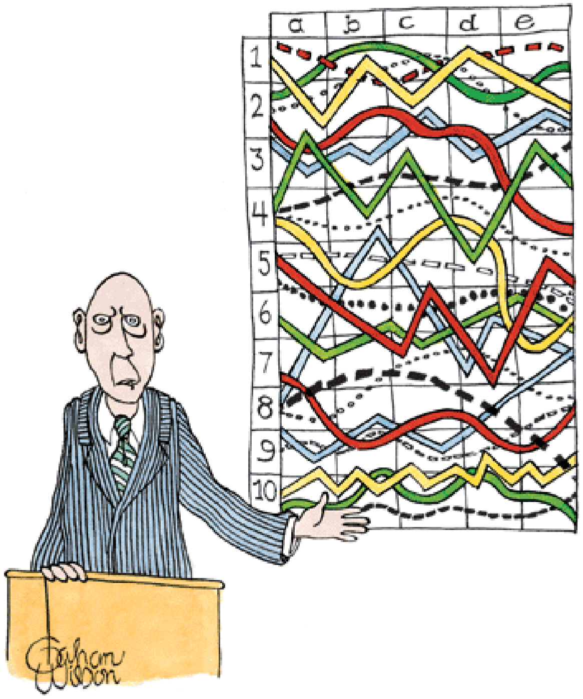
Is there value in an international funding mechanism?

User Community

- Who are the users by discipline?
- How many?
- How do you know?

Data Usage

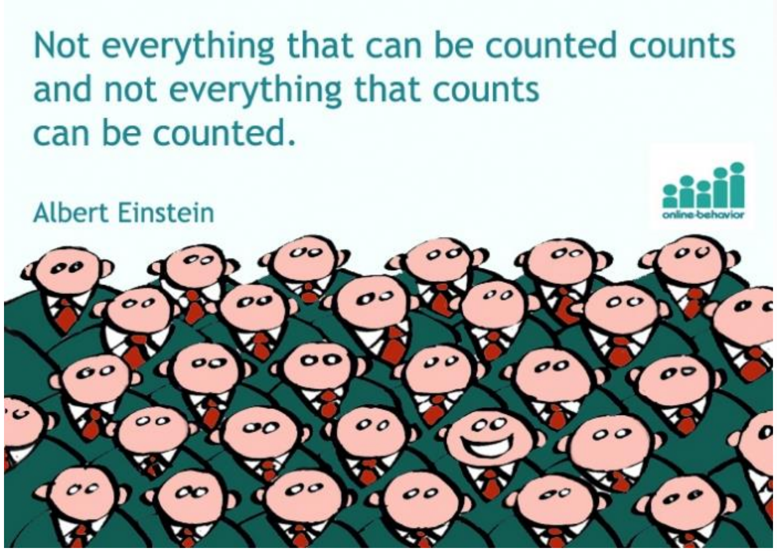
- What are the data used for?
- How do you monitor data usage?
  - Web access
  - Download statistics
  - Citations to data
  - Citations to data resource



"I'll pause for a moment so you can let this information sink in."

Measuring Impact

- Metrics
  - Web usage
  - Database IDs listed in literature
  - Data reuse
- Effect on research
  - Enables new research
  - Time saved


Not everything that can be counted counts
and not everything that counts
can be counted.
Albert Einstein
online-behavior

What If Your Resource Ceased to Exist?

- Good and bad disruption
- What would be hindered?
  - Reproducibility
  - Experimental design
  - Publication pipeline
  - ?
- Consequences of establishing a new resource
  - Financial
  - ”Corporate History”

Contingency Plans

- Some of the responses
  - None
  - Has happened already
  - Multiple funders provide a backup

Challenges

- Funding
  - Current funding and review mechanisms are not appropriate for infrastucture
  - Short duration of grants
  - Agencies like creation but not renewal of resources
- Lack of understanding of importance of curation

Data Resources Being Presented Today

Data Resources for the Life Sciences		**Chair:** **Helen Berman**
0945	Introduction to Data Resources	Helen Berman
1000	The Universal Protein Resource (UNIPROT)	Alex Bateman
1015	The Worldwide Protein Databank (wwPDB)	Stephen Burley
1030	The Online Mendelian Inheritance in Man (OMIM)	Ada Hamosh
1045	The Proteome Xchange Consortium	Henning Hermjakob
1100	**Coffee Break**	
1115	The Kyoto Encyclopedia of Genes and Genomes (KEGG)	Minoru Kanehisa
1130	The International Nucleotide Sequence Database Collaboration (INSDC)	Guy Cochrane
Model Organism Databases		
1145	Alliance of Genome Resources: Model Organism Databases (MODs) join forces	Paul Sternberg
1200	Mouse Genome	Judith Blake
1215	Flybase	Norbert Perrimon
1230	Zfin	Monte Westerfield
1245	Summary of major issues for data resources/model organisms and open discussion	