Personal genomes and privacy: perspectives from a public repository

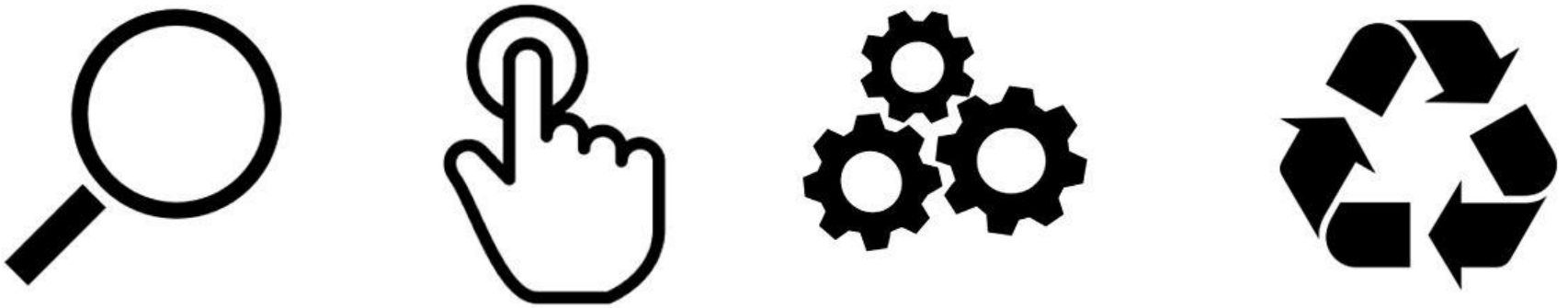
Masanori Arita

Bioinformation & DDBJ Center
National Institute of Genetics

arita@nig.ac.jp

9th Big Data Forum in Beijing, China

Findable Accessible Interoperable Reusable



FAIR principle is well known to achieve research integrity and equity.

But it is not fully compatible with international fora!

Access to DSI or digital sequence information now invokes “benefit sharing”, just like access to genetic materials.

Challenges

International treaties are **NOT** harmonized in terms of data sharing policy and management.

Treaty	CBD (biodiversity)	BBNJ (high seas)	WHO Pandemic (covid-19 etc)	ITPGRFA (crops in FAO)
Target information	DSI (not yet clear)	DSI (not yet clear)	Human and virus genomes	Crop genomes
Repository or Indexing site	Clearing House in each nation	Global clearing house	Hospitals, GISAID	Global Information System (GLIS)
Open Access / FAIR data	**Nation-wise**	**YES**	**NO**	**?**
Benefit Sharing mechanism	Under discussion	multilateral	No benefit; COVAX?	multilateral
Related Documents	https://www.cbd.int/abs/	IISD Bulletin: bit.ly/bbnj5res	UN Doc SSA2/CONF./1Rev.1	CBD.INT ITPGRFA-DSI.pdf

Why cannot we simply go for “open”?

Nowadays, control-rights or ownership is at stake even for scientific information (esp. for IPLCs).

From: Localcontexts.org

> UNDRIP affirms Indigenous Peoples’ rights and interests in their data. Recognition of these rights bolsters Indigenous Peoples’ **authority to control** and govern such data, further affirming the need for ‘data for governance.’ Indigenous Peoples must have access to data that support Indigenous governance and self-determination.

Carroll et al. *Data Science J.* DOI: 10.5334/dsj-2020-0

Data control is not easy.

COVID-19 pandemic told us important examples to consider.

1. Travel ban to South Africa who quickly reported the Omicron variant
2. Only 0.3% vaccine distribution to LMICs (when 30% is disposed)
3. Disinformation on vaccines and other remedies

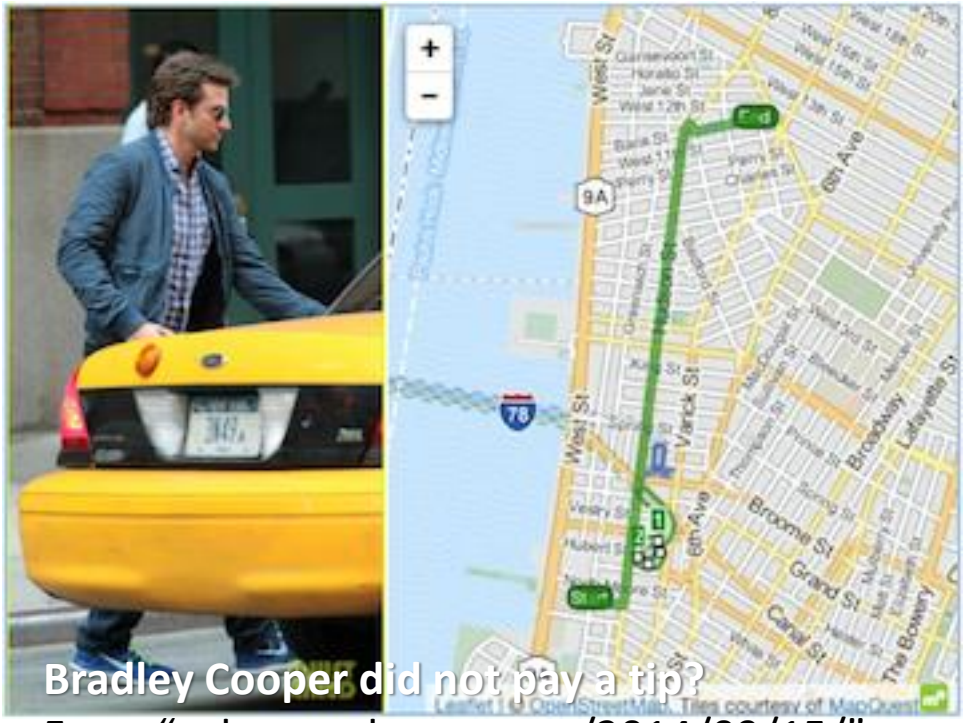
Lessons learned: Incentivize and credit data depositors. Enable high-quality metadata. *Intelligent* openness!

Just opening data may cause problems:
Passenger Privacy in the NYC Taxicab Dataset

Anonymized taxi data set was immediately hacked with gossip blogs to reveal:

- Trip and tip amount of celebrities
- Addresses who frequently visit a famous gentlemen's club

:



From "agkn.wordpress.com/2014/09/15/"

Important lessons:

Value cannot be determined by the data alone.

Hacking is not recommended but we can learn a lot!

What we do as a public repository?

Tools to analyze “open” data from 1000 Genomes.

- Togo-Imputation server

- Use conform-gt (version 24May16) to convert the reference / alternative allele of the input SNP array data to match the reference panel data.
- Use Beagle 5.2 (version 21Apr21.304) for fading and imputation analysis.
- Index the genomic data (VCF file) after imputation using bcftools (version 1.9).

A series of workflows are implemented in Common Workflow Language (CWL) and published as an input-server workflow.

- Parabricks & other pipelines for GRCh37, 38, and other references (please visit ToMMo).

Benefit-sharing from sequences

In COP15 (Kunming-Montreal, Dec 2022) and COP16 (Columbia, Oct 2024), benefit-sharing scheme for DSI is under discussion.

- Human sequences are out of scope.
  - How about pathogens? Gut microbiome? Metabolome?
- Academic databases are not exempted.
  - How do we assess “benefit” / “profit” / “income” ?

Please see our position paper of “DSI Scientific Network”:
https://www.dsiscientificnetwork.org/resources/

Basic standpoints of academia

These are not “common sense”, we need to **defend** our position.

Commercial sectors may ask academia to contribute money to reduce *their responsibility*.

1. **Non-commercial users should contribute to non-monetary benefits, not to the Fund**

2. **DSI capacity building should be a focus area for disbursement of funds**

3. **Further work is needed on Non-Monetary Benefit-Sharing**

4. **Advancing benefit sharing does not require a new database under the CBD**

5. **Database practices can be improved, but database managers cannot be the ABS police**

COP16 Position paper of “DSI Scientific Network”:
https://www.dsiscientificnetwork.org/resources/

“Privacy” needs re-definition

A diverse concept defining the relationship between social entities:

- Context-dependent (what is exposed, how, and when by whom?)
- Critical for individual well-being
- Often paradoxical (few people pay attention)

Loss of contextual integrity will lead to privacy harms, but how can we prevent harms?

Data Ethics Policy Briefs
https://codata.org/initiatives/task-groups/data-ethics/

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Reparation scheme is needed

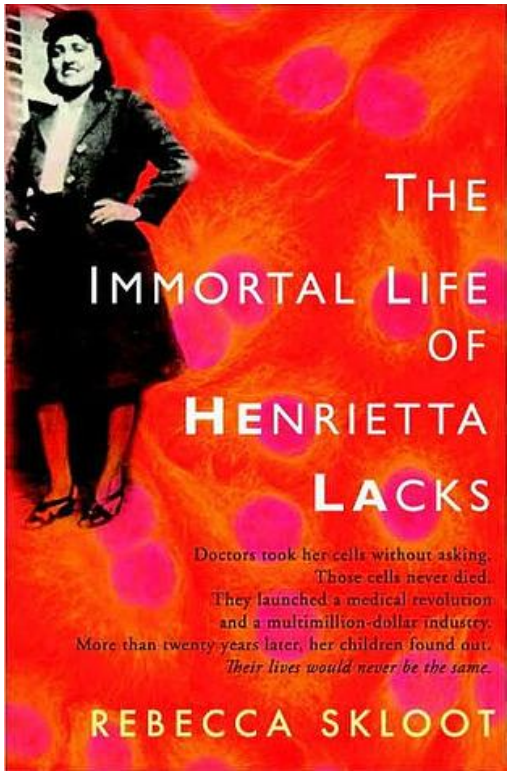
HeLa cells from Henrietta Lacks (1951) contributed to science significantly.

- Polio & papillomavirus vaccines
- basic science

The value of specimen and genomes will change with time and technology.

HeLa History

A single book in 2010 ignited discussion on the reparations for unfair exploitation of HeLa cells. Now several pharma have compensated with her families.



Conclusions

- Please watch out for the coming COP16 on CBD.
- We need to clearly define what is “personal” in genome science. (GDPR is another viewpoint.)
- Reparation / compensation is important.
- We need to **unite** to sustain the current FAIR data scheme.

Important Links

Convention on Biological Diversity, Digital Sequence Information website

https://www.cbd.int/dsi-gr/whatdone.shtml

DSI Scientific Network policy brief

https://www.dsiscientificnetwork.org/resources/

CODATA Data Ethics policy brief

https://codata.org/initiatives/task-groups/data-ethics/