Collaboration of ELIXIR and Japan BioHackathons

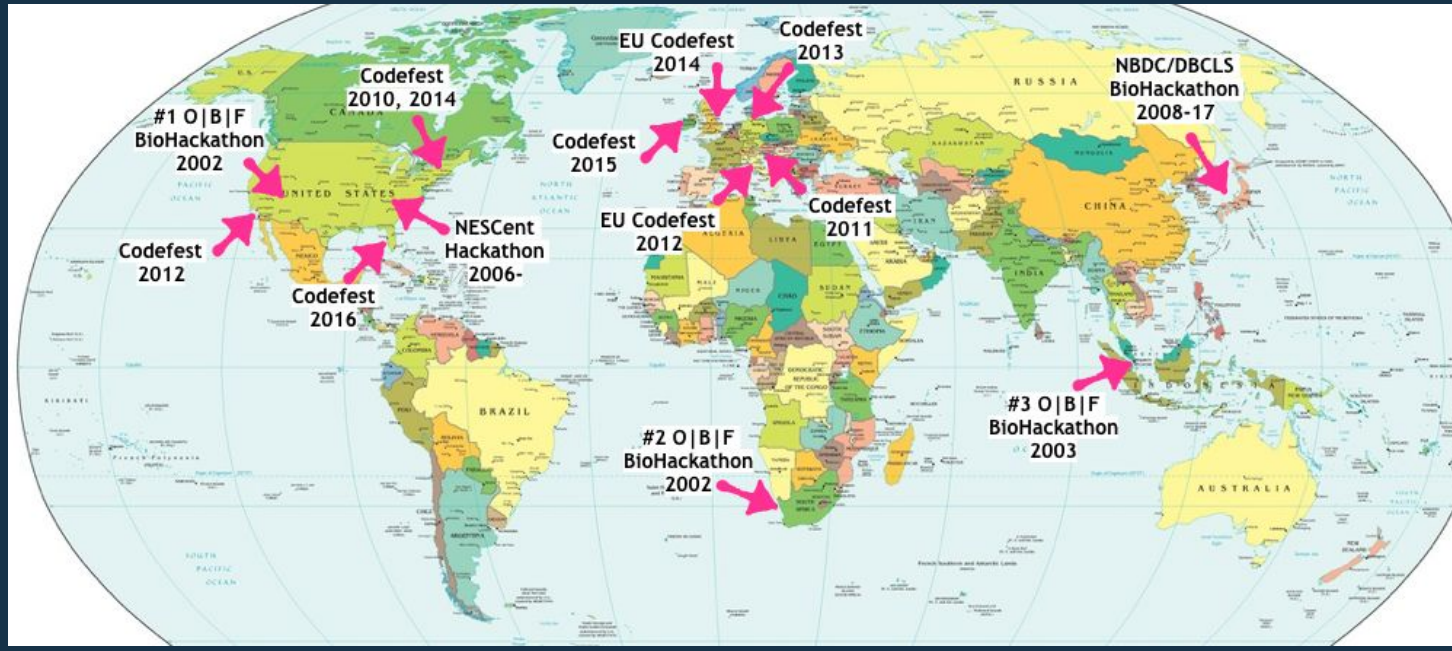
Toshiaki Katayama
- Database Center for Life Science

Project links
- http://biohackathon.org/
- http://github.com/dbcls/bh18/

Background information

How we get started the BioHackathon?

- BioHackathon was originally introduced by the O|B|F in 2002



- BioPerl
- BioRuby
- Biopython
- BioJava

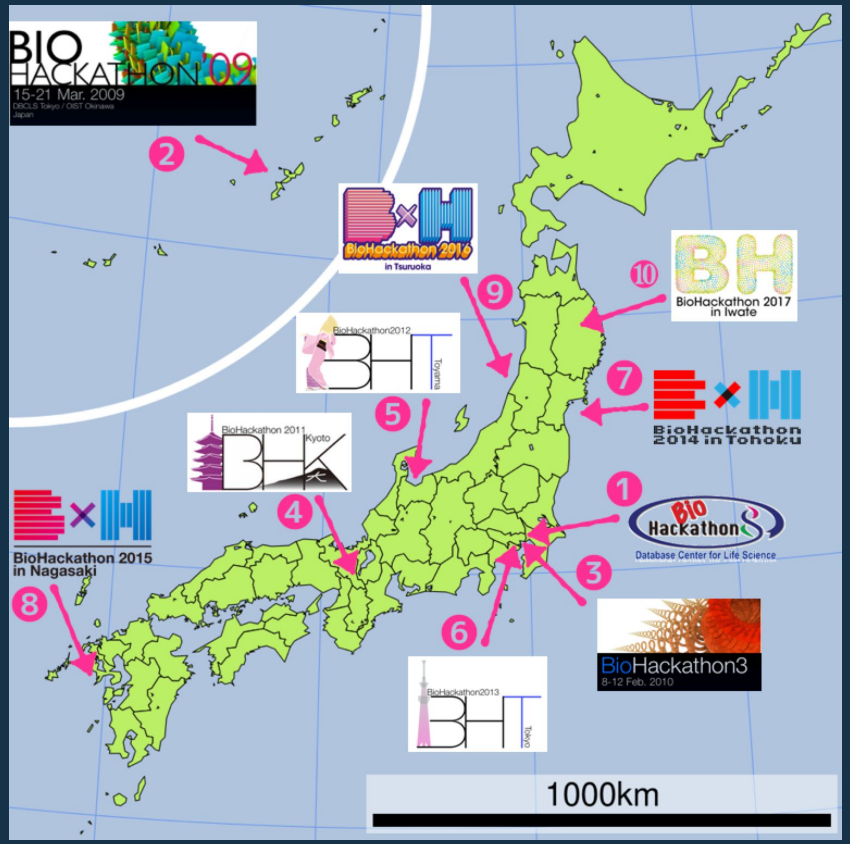
  :

- Interoperability of Open Bio* libraries for accessing DBs

History of the NBDC/DBCLS BioHackathon

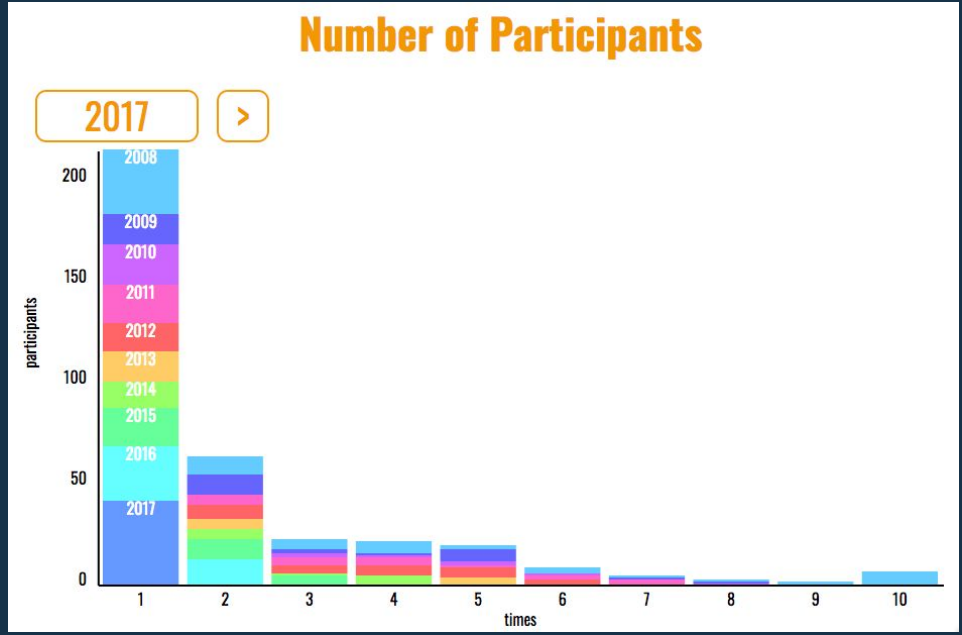
- BioHackathon 2008-2010
  - Integration of distributed bioinformatics resources with Web Services
- BioHackathon 2011-2017
  - Integration of distributed bioinformatics database contents with Semantic Web

- As we were not secured to organize the next BioHackathon after 10th, I explored possibility to organize it with ELIXIR (especially with Rafael Jimenez, thanks!)
  - → This ELIXIR BioHackathon 2018
  - ＋ NBDC/DBCLS BioHackathon 2018



The 11th NBDC/DBCLS BioHackathon

- NBDC/DBCLS BioHackathon 2018
  - Application of integrated semantic resources including (but not limited to)
    - 1. Biomedical
    - 2. Useful substance production
    - 3. Breeding
- Register today (already over due)!
  - December 9-15 in Matsue, Japan
  - http://2018.biohackathon.org/registration
  - We'll welcome Matz (the Ruby's father)



Each year, we had 68 (2008) ~ 108 (2017) participants, in total, 803 participants (371 uniq) from 21 countries.

http://biohackathon.org/10years/

Hack organisation

Organisation of the NBDC/DBCLS BioHackathon

- Attendees
  - Call for proposals to be nominated as an invitee
  - 20~30 foreign experts and invitees + 50~80 Japanese domestic researchers (incl. organizers)
  - Can form working groups and switch to other group freely
- Schedule
  - Sunday: Public symposium
  - Monday-Friday: Hackathon
    - Monday morning: Open space discussion to define objectives and working groups
    - Wednesday: Mid-term wrap up (Excursion sometimes)
    - Friday afternoon: Wrap up
  - Saturday: Writethon
- Venue
  - 1Gb (x2 for backup) fibre optical Internet connection
  - (24hrs open) hacking room + accommodation
  - Isolated place hopefully equipped with Onsen (hot spring) and Karaoke
  - Fine food and Sake (Liquorthon at night)

Powered by BioHackathon

Previous publications

1. Katayama, T., Arakawa, K., Nakao, M., et al. (2010) **The DBCLS BioHackathon**: standardization and interoperability for bioinformatics web services and workflows., J. Biomed. Semantics, 1, 8.
2. Katayama, T., Wilkinson, M. D., Vos, R., et al. (2011) **The 2nd DBCLS BioHackathon**: interoperable bioinformatics Web services for integrated applications., J. Biomed. Semantics, 2, 4.
3. Katayama, T., Wilkinson, M. D., Micklem, G., et al. (2013) **The 3rd DBCLS BioHackathon**: improving life science data integration with semantic Web technologies., J. Biomed. Semantics, 4, 6.
4. Bolleman, J. T., Mungall, C. J., Strozzi, F., et al. (2016) **FALDO**: a semantic standard for describing the location of nucleotide and protein feature annotation., J. Biomed. Semantics, 7, 39.
5. Katayama, T., Wilkinson, M. D., Aoki-Kinoshita, K. F., et al. (2014) **BioHackathon series in 2011 and 2012**: penetration of ontology and linked data in life science domains., J. Biomed. Semantics, 5, 5.
6. Wimalaratne, S. M., Bolleman, J., Juty, N., et al. (2015) SPARQL-enabled identifier conversion with **Identifiers.org**., Bioinformatics, 31, 1875–7.
7. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., et al. (2016) **The FAIR Guiding Principles** for scientific data management and stewardship, Sci. Data, 3, 160018.
8. ...
9. BioHackathon 2013-2014 (submitted)
10. BioHackathon 2015 (revise submitted)
11. BioHackathon 2016 (in preparation)
12. BioHackathon 2017 (in preparation)

Powered by BioHackathon

Future works

- Continuous community efforts on standardization and interoperability of databases in life sciences
  - Genomics and Multi-omics
  - Biomedical applications
  - Bioinformatics algorithms and analysis
  - Reproducible science and workflows
- Collaboration with biomedical domains
  - GA4GH (Global Alliance for Genomics and Health
    - Bring algorithms to data: WES, CWL, ...
    - Sharing and standard models for genotypes and variations
    - Graph genome: vg
  - AMED (Japan agency for medical research and development)
    - Japanese gnomAD and ClinVar: TogoVar, MGeND
  - and ELIXIR, of course!

Some photos from BioHackathon 2017

Standardization matters

- NIH strategic plan for data science
  - https://datascience.nih.gov/strategicplanrelease
  - According to a 2016 survey, data scientists across a wide array of fields said they spend most of their work time (about 80 percent) doing what they least like to do: collecting existing datasets and organizing data. That leaves less than 20 percent of their time for creative tasks like mining data for patterns that lead to new research discoveries.

- Imagine if
  - Data you need is already stored in an uniformed format
  - All these heterogeneous data are interconnected with globally unique identifiers
  - And semantically annotated!
    - Linked Open Data (see UniProt as one of successful examples)
  - You can just start data science and develop applications

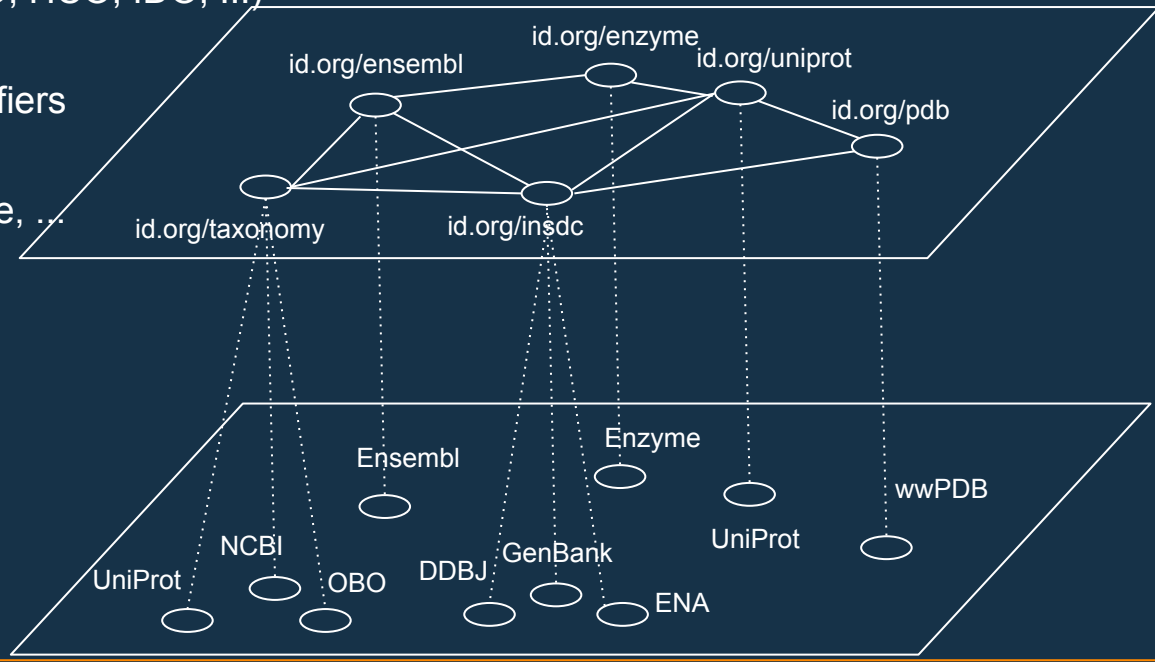
RDF resources currently available

- Nucleotide seq & annot
  - INSDC (DDBJ/DBCLS)
- Genome
  - Ensembl (EBI)
  - RefSeq (TogoGenome)
- Protein seq & annot
  - UniProt (SIB)
- Protein structure
  - PDB (PDBj)
  - BMRB (PDBj)
  - FAMSBASE (Chuo U)
- Compounds
  - PubChem (NCBI)
  - ChEMBL (EBI)
  - Nikkaji (JST)
- Gene expression
  - RefEx (DBCLS)
  - ExpressionAtlas (EBI)
- Samples
  - BioSamples (EBI/DDBJ)
  - JCM (RIKEN)

- Biomedical (Med2RDF)
  - ICGC, COSMIC, CIViC
  - DGIdb, OpenTG-Gates
  - ClinVar, dbSNP, dbVar
  - ExAC, gnomAD
  - HiNT, INstruct
- Glycome
  - GlyTouCan, GlycoEpitope, WURCS, GGDonto, PAConto
- Proteome
  - jPOST
  - The Human Protein Atlas
- Pathway
  - Reactome (EBI)
- Others
  - MeSH (NCBI)
  - BioModels (EBI)
  - MBGD (NIBB/DBCLS)
  - Quanto (DBCLS)
  - :

Standardization requires community effort

- Data models
  - Organism - Genome - Gene - Variation - Phenotype - Disease - Drug - Multi-omics - ...
- Ontology
  - BioPortal and others (incl. FALDO, HCO, IDO, ...)
- Identifiers
  - Identifiers.org and compact identifiers
- Metrics
  - FAIR, YummyData, RDF guideline, ...
- Workflows
  - WES, CWL, Galaxy, ...
- Reusable Web Components
  - BioJS, TogoStanza, ...



Thank you and join us!

- Organizing a hackathon is a demanding task but should be fruitful for the future
  - Many thanks and congratulations to the ELIXIR BioHackathon organizers especially to Rafa, Dana, Layla and local organizers
  - Let's get started and have fun!

- And hopefully see you next month at the NBDC/DBCLS BioHackathon 2018
  - Register NOW! http://2018.biohackathon.org/

Will try to be a better thinker :)

I really appreciate organizers for the surprise gift!