Standardization of biological sample information database

Tazro Ohta1*, Shuya Ikeda1, Takatomo Fujisawa2, Shuichi Kawashima1

1 Database Center for Life Science (DBCLS), ROIS-DS

2 DNA Data Bank of Japan (DDBJ), National Institute of Genetics, ROIS

Email: t.ohta@dbcls.rois.ac.jp

International Symposium "Global Collaboration on Data beyond Disciplines"

Early Career Researcher (ECR) session

25 September 2020

Me

1. Name: Tazro Ohta @inutano

2. Affiliation: DBCLS, ROIS-DS

3. Position: Project Assistant Professor

4. Years from Ph.D. received: 1.5y (March 2019)

5. What kind of data you are handling now: Genomics (Genes, Genomes, Cells, etc.)

6. Category of your research: Data and Database research and development

7. Research, or job, position you would like to do or become in the future: Open source community researcher

8. Your photo: :D

Agenda

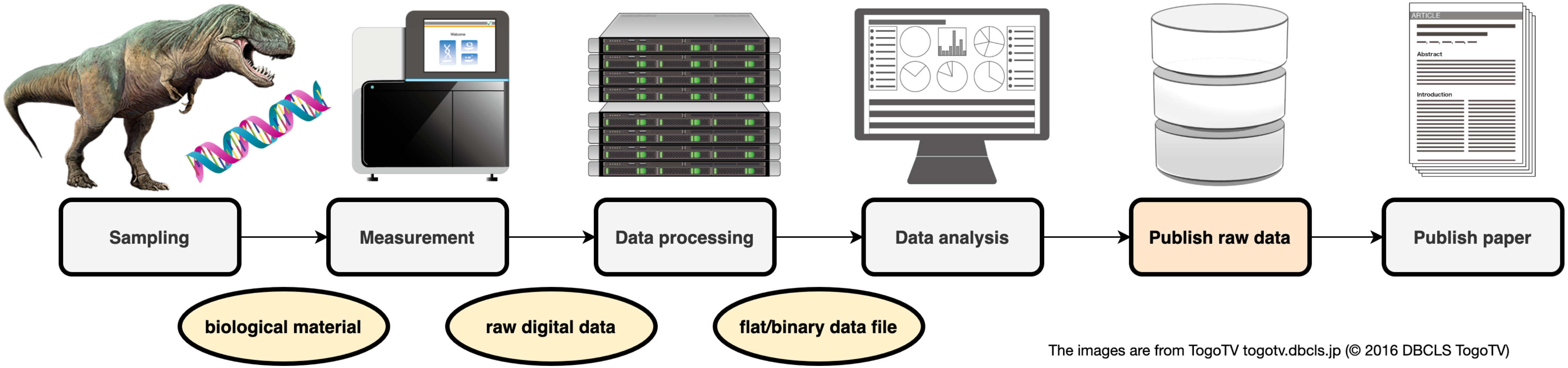
- Background:
    - Sharing genomic data across the biomedical community
    - Building knowledge upon the public database
- Problem:
    - Describing biological sample metadata
- Solution: 2 approaches
    - a. Ontology mapping
    - b. Data modeling with RDF

Background

Sharing genomic data across the biomedical community

- The Bermuda Principles (1996)
  - Rules for publishing DNA sequence data
  - Publish DNA sequence data before publishing paper
  - Public domain license to research usage
- International Nucleotide Sequence Database Collaboration (INSDC)
  - NCBI (US), EBI (EU), DDBJ (Japan)
  - Exchange submissions (mirroring)
  - Archiving over 44P bases

Publish DNA sequence data BEFORE publishing paper



Sampling → Measurement → Data processing → Data analysis → **Publish raw data** → Publish paper

biological material

raw digital data

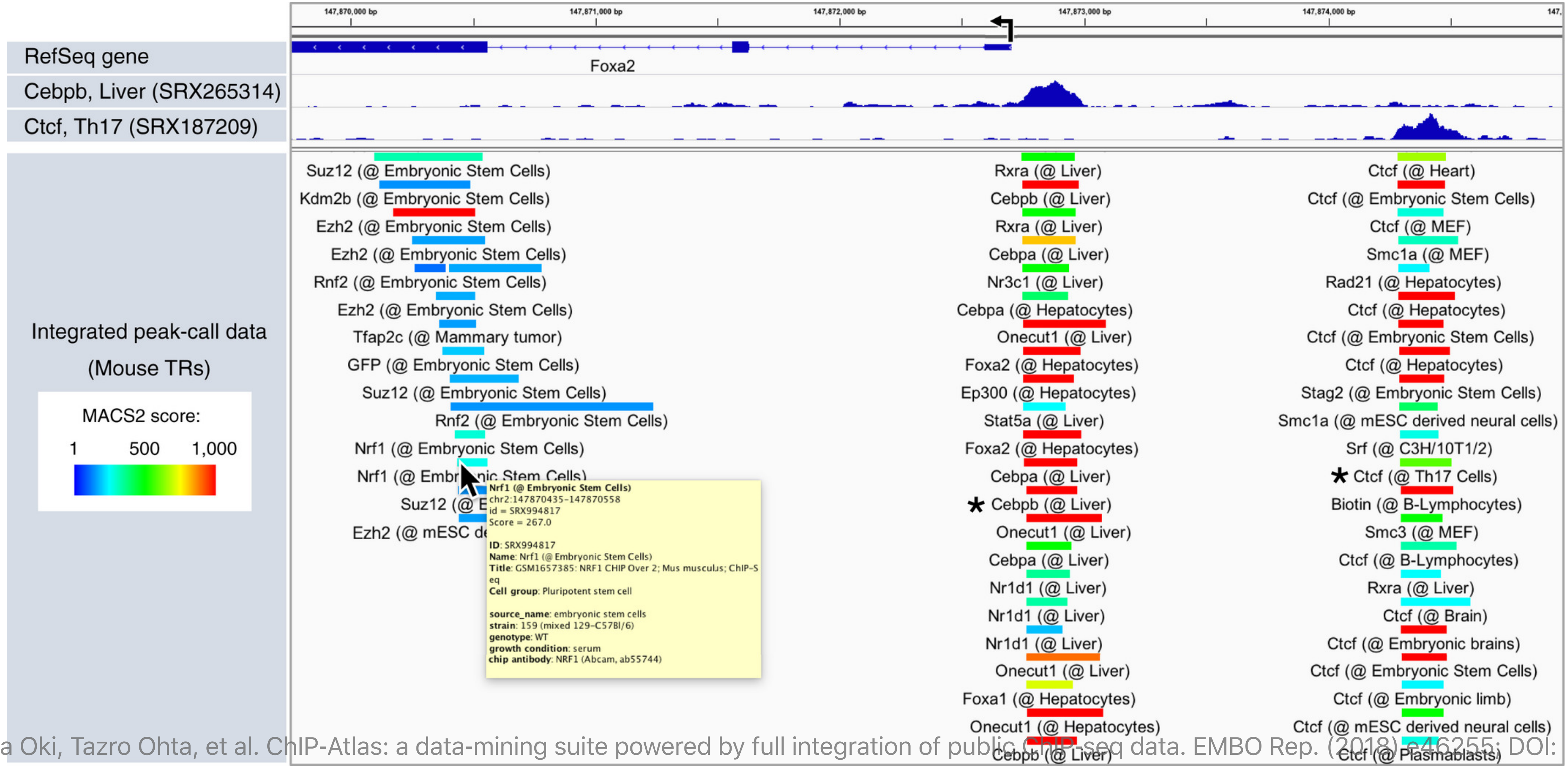
flat/binary data file

The images are from TogoTV togotv.dbcls.jp (© 2016 DBCLS TogoTV)

Building knowledge upon the public database

- Public domain: free to use for research purpose
- International consortiums provide comprehensive data for a specific domain
    - 1000 genomes project
    - ENCODE project
- Researchers build secondary databases based on public data
    - ChIP-Atlas

ChIP-Atlas: process all the public ChIP-seq data



Nrf1 (@ Embryonic Stem Cells)
chr2:147870435-147870558
id = SRX994817
Score = 267.0

ID: SRX994817
Name: Nrf1 (@ Embryonic Stem Cells)
Title: GSM1657385: NRF1 CHIP Over 2; Mus musculus; ChIP-Seq
Cell group: Pluripotent stem cell

source_name: embryonic stem cells
strain: 159 (mixed 129-C57Bl/6)
genotype: WT
growth condition: serum
chip antibody: NRF1 (Abcam, ab55744)

Shinya Oki, Tazro Ohta, et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Rep. (2018) e46255; DOI: https://doi.org/10.15252/embr.201846255

Problem

Describing biological sample metadata

- BioSample
    - Public database archiving information of biological material used in experiments
    - Submitters describe key-value pairs to explain single biological material
    - over 8M samples and growing
- **Problem:** inconsistent sample description
    - Different keys for the same concept
    - Different form of same values
    - Synonyms
    - typos
- How can we handle those variations?

BioSample record examples

```
tissue: blood
age: 45
sex: male
```

```
tissue: Blood
age: 45 years old
biological_sex: M
```

```
cell type: HeLa
```

```
cell_line: HeLa
```

```
cell type: adipocyte
```

```
cell type: fat cell
```

Solution: 2 approaches

1. Ontology mapping

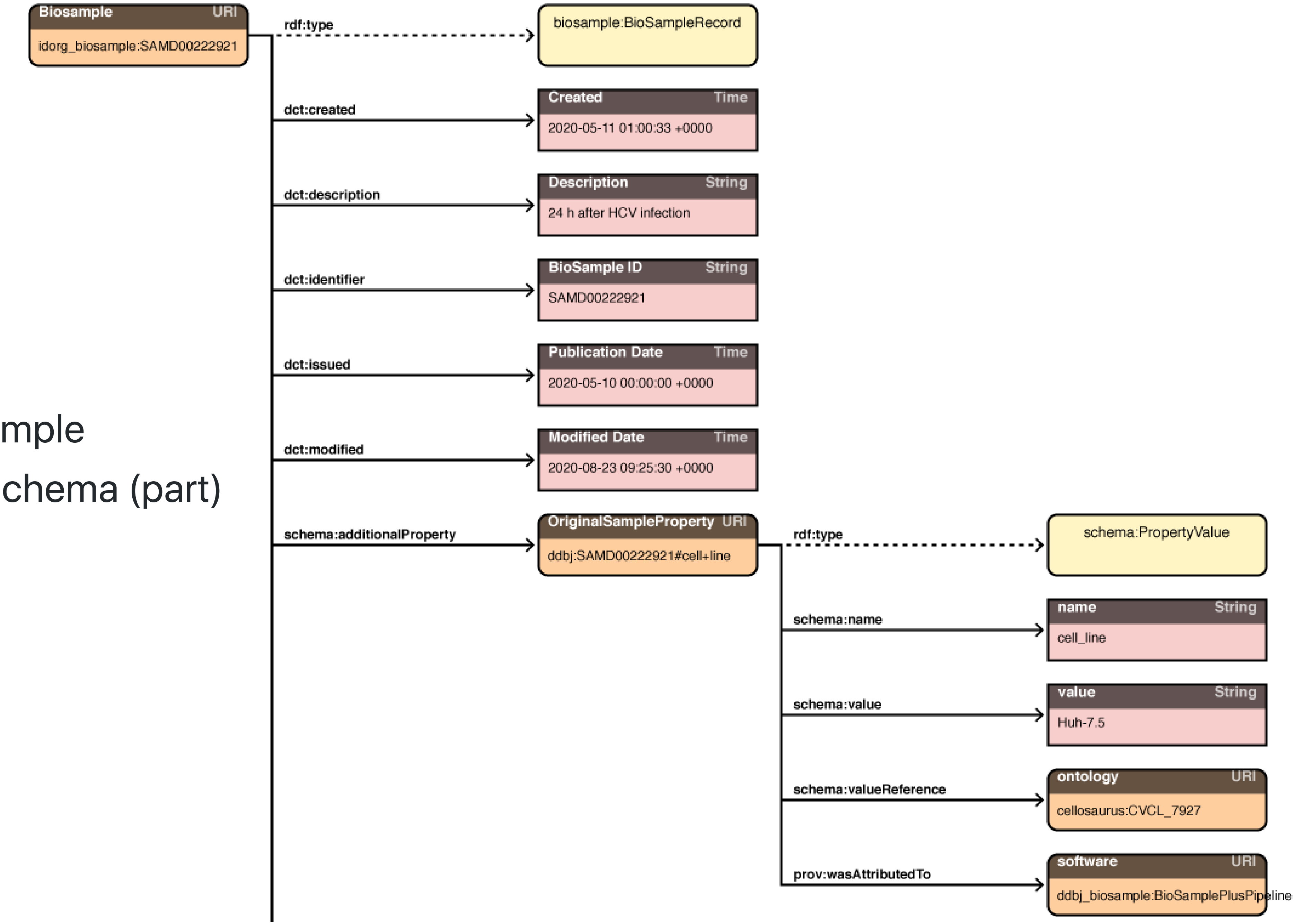
- MetaSRA: an existing software/database to map ontology to sample description
    - Only for a specific type of experiment
- Improved MetaSRA implementation for
    - faster execution (6h to 1h for 5000 samples)
    - ontology term optimization

2. Data modeling with RDF

- RDF: Resource Description Framework
    - A W3C standard for data description
    - Using URI to identify resources, linking things
- Why RDF?
    - Interoperability: suitable for biological data: many different small domains
        - genes, proteins, diseases, etc.
    - Many biological databases are now provided in RDF form
        - https://integbio.jp/rdf

BioSample

RDF schema (part)



Biosample URI
idorg_biosample:SAMD00222921

rdf:type
biosample:BioSampleRecord

dct:created
Created Time
2020-05-11 01:00:33 +0000

dct:description
Description String
24 h after HCV infection

dct:identifier
BioSample ID String
SAMD00222921

dct:issued
Publication Date Time
2020-05-10 00:00:00 +0000

dct:modified
Modified Date Time
2020-08-23 09:25:30 +0000

schema:additionalProperty
OriginalSampleProperty URI
ddbj:SAMD00222921#cell+line

rdf:type
schema:PropertyValue

schema:name
name String
cell_line

schema:value
value String
Huh-7.5

schema:valueReference
ontology URI
cellosaurus:CVCL_7927

prov:wasAttributedTo
software URI
ddbj_biosample:BioSamplePlusPipeline

Provide BioSample RDF data to the community

- ftp://ftp.ddbj.nig.ac.jp/rdf/biosample/
- Mapped ontology aligned same concepts to a single value
- RDFized BioSample data can link to the external database records with no extra effort

Summary

- Sharing genomic data has advanced the whole biomedical research
- Describing explicit biological sample metadata is essential for data reuse
- Graph-based data model with mapped ontology terms helps better data sharing