Metadata Reconciliation
for Genomic Datasets

Anna Bernasconi, PhD student
Arif Canakoglu, Postdoc researcher
Andrea Colombo, Master Graduate

Challenges in Data-Driven Genomic Computing Workshop – March 6-8th 2019
Organized by "GeCo" – Data-Driven Genomic Computing

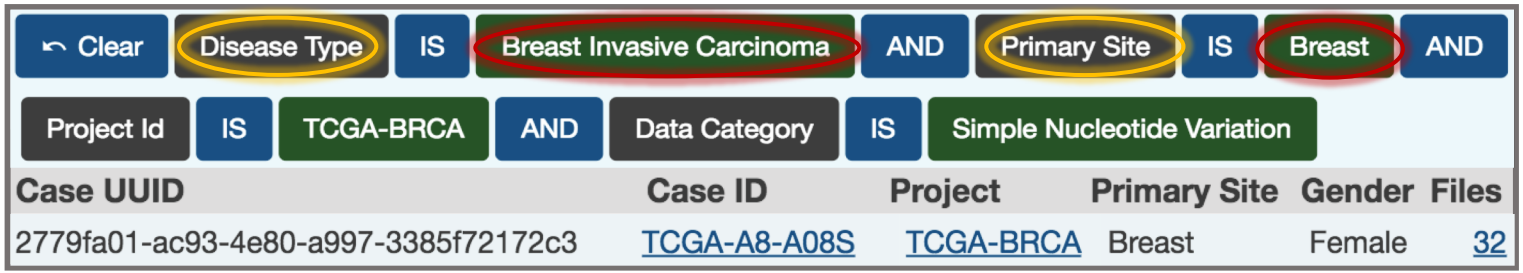
POLITECNICO
MILANO 1863

Motivation for metadata curation in Genomics

- A lot of emphasis on **data production** and **sharing** 
- Some attention on genomics **data quality** 
- Poor accuracy of **data documentation** (i.e., metadata): 
  - Not standardized across sources
  - Unstructured
  - Incomplete

Lack of:
- Unifying metadata conceptual model
- Structured integration procedures
- Interoperability

Genomic Data Commons



Gene Expression Omnibus

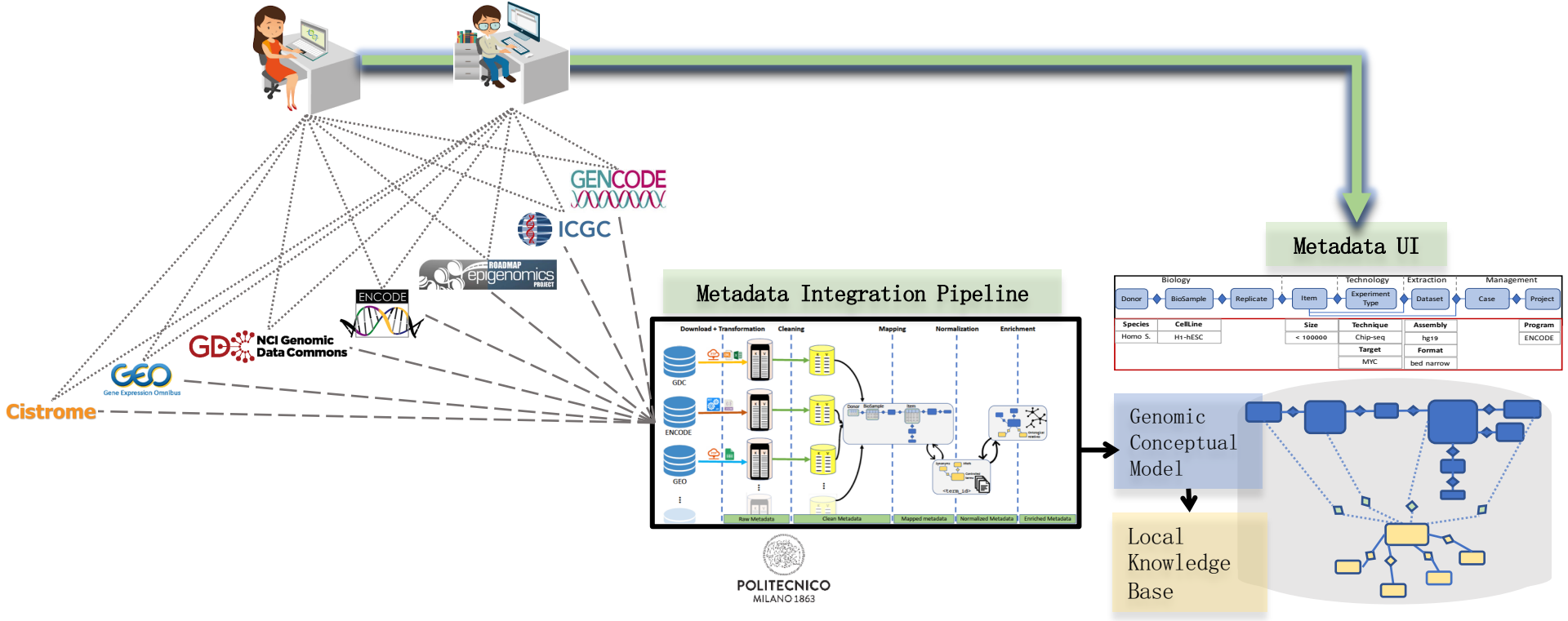
Sample GSM1197482 Query DataSets for GSM1197482

Source name	T47D-MTVL
Organism	Homo Sapiens
Characteristics	gender: female
	tissue: breast cancer ductal carcinoma

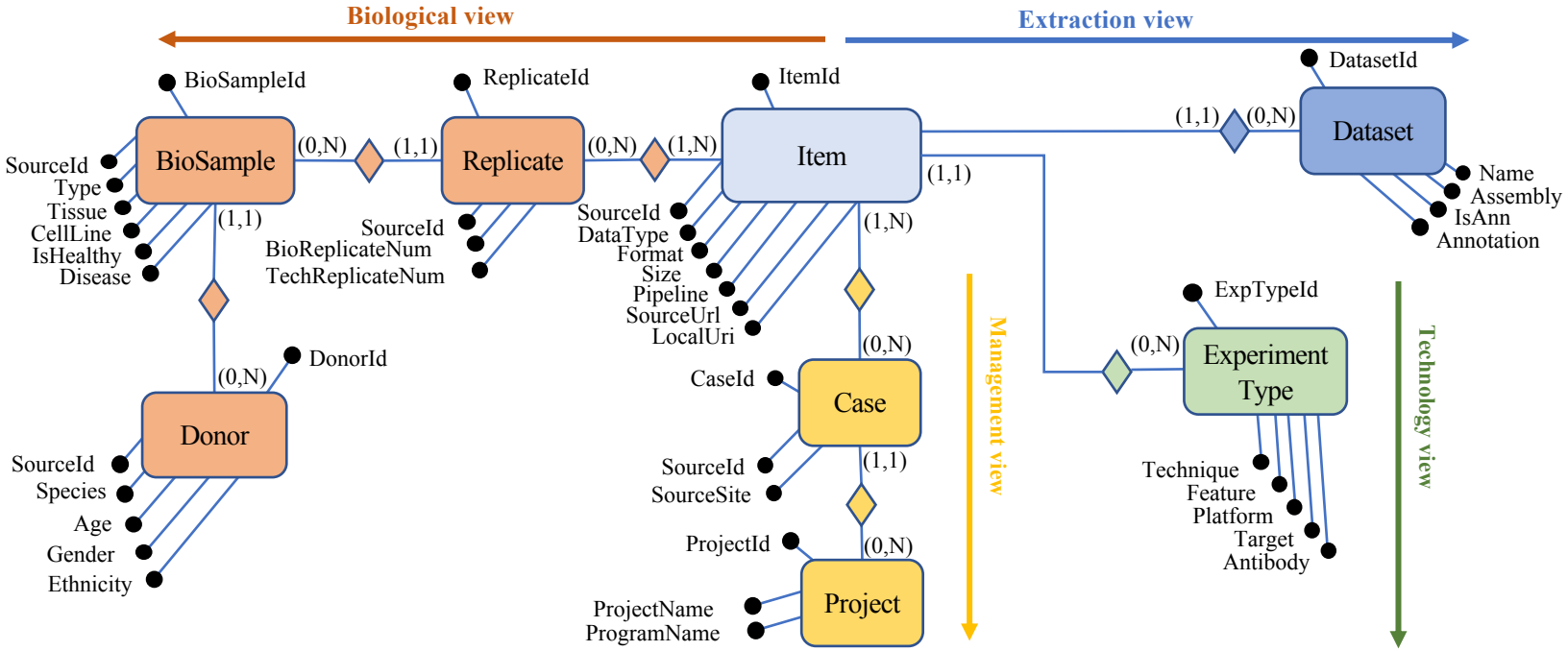
ENCODE

Experiment summary for ENCSR000DMQ		**Experiment summary for ENCSR000DOS**	
Assay:	ChIP-seq	**Assay:**	ChIP-seq
Target:	MYC	**Target:**	MYC
Biosample:	*Homo sapiens* MCF-7	**Biosample:**	*Homo sapiens* MCF-10A
Biosample Type:	cell line	**Biosample Type:**	cell line
Description:	Mammary gland, adenocarcinoma	**Description:**	Mammary gland, non-tumorigenic cell line
Health status:	Breast cancer (adenocarcinoma)	**Health status:**	Fibrocystic disease

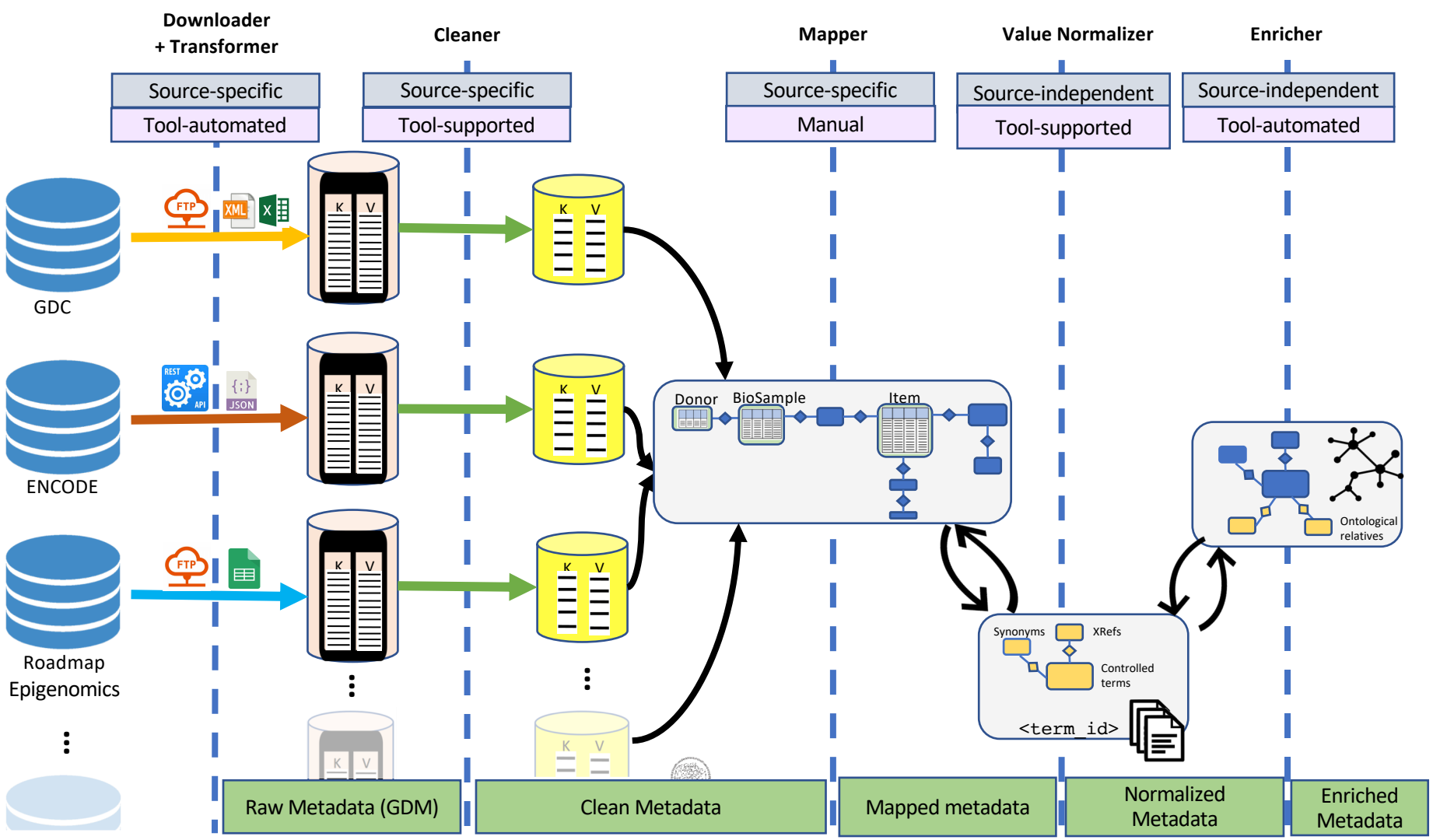
Our approach



The Genomic Conceptual Model (GCM)



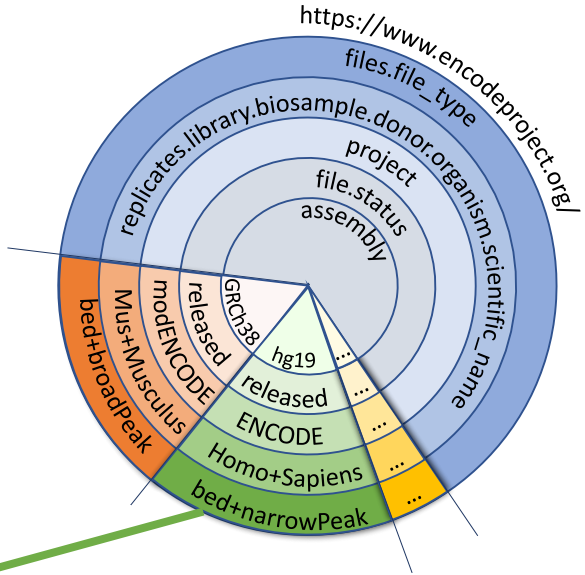
Bernasconi A., Ceri S., Campi A., Masseroli M. (2017)
Conceptual Modeling for Genomics: Building an Integrated Repository of Open Data, ER2017.


Downloader + Transformer
Cleaner
Mapper
Value Normalizer
Enricher
Source-specific
Tool-automated
Source-specific
Tool-supported
Source-specific
Manual
Source-independent
Tool-supported
Source-independent
Tool-automated
GDC
ENCODE
Roadmap Epigenomics
FTP
XML
REST API
JSON
K
V
Donor
BioSample
Item
Ontological relatives
Synonyms
XRefs
Controlled terms
<term_id>
Raw Metadata (GDM)
Clean Metadata
Mapped metadata
Normalized Metadata
Enriched Metadata

Downloader

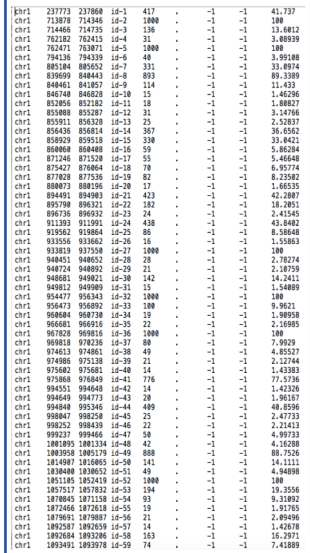
https://www.encodeproject.org/metadata/?type=Experiment&⟨params⟩/metadata.tsv
https://www.encodeproject.org/files/⟨file_accession⟩/@@download/⟨file_accession⟩.bed.gz

Encode
assembly = hg19
file.status = released
project = ENCODE
replicates.library.biosample.donor.organism
.scientific_name = Homo+sapiens
files.file_type = bed+narrowPeak



Transformer

Downloaded region file

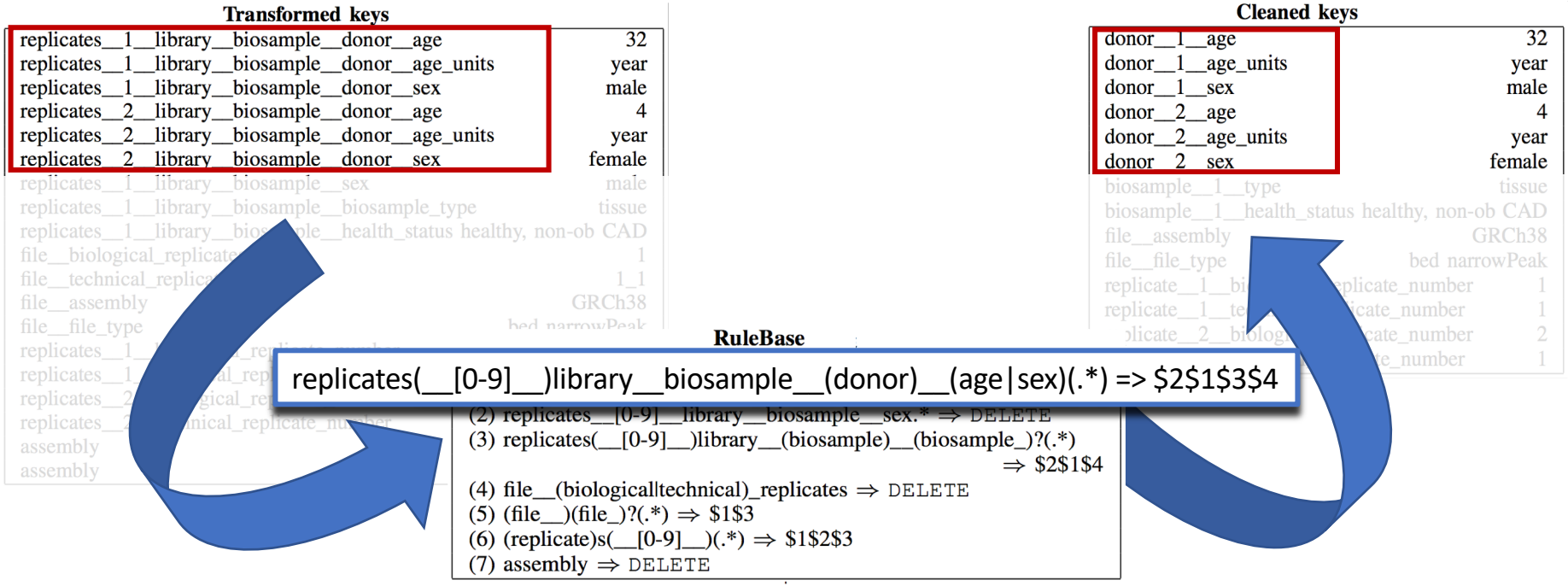


Downloaded metadata file

```
{"accession": "ENCSR635OSG",
 "files": [
     {"accession": "ENCFF002EWY",
      "biological_replicates": [1],...},
     {"accession": "ENCFF429VMY",
      "biological_replicates": [1,2],
      "file_type": "bed narrowPeak",...},
      ...
 ],
 "replicates": [
     {
         "@id": "/replicates/4874c170-7124-4822-a058-4bb/",
         "biological_replicate_number": 1,
         "library": {
             "biosample": {
                 "donor": {"age": "6",...},
                 "health_status": "healthy",
                 ...},
             ...},
         ...
     },
     {
         "@id": "/replicates/d42ff80d-67fd-45ee-9159-25a/",
         "biological_replicate_number": 2,
         "library": {
             "biosample": {
                 "donor": {"age": "32",...},
                 "health_status": "healthy with non-
                     obstructive coronary artery disease",
                 ...},
             ...},
         ...
     }
 ],
 ...}
```

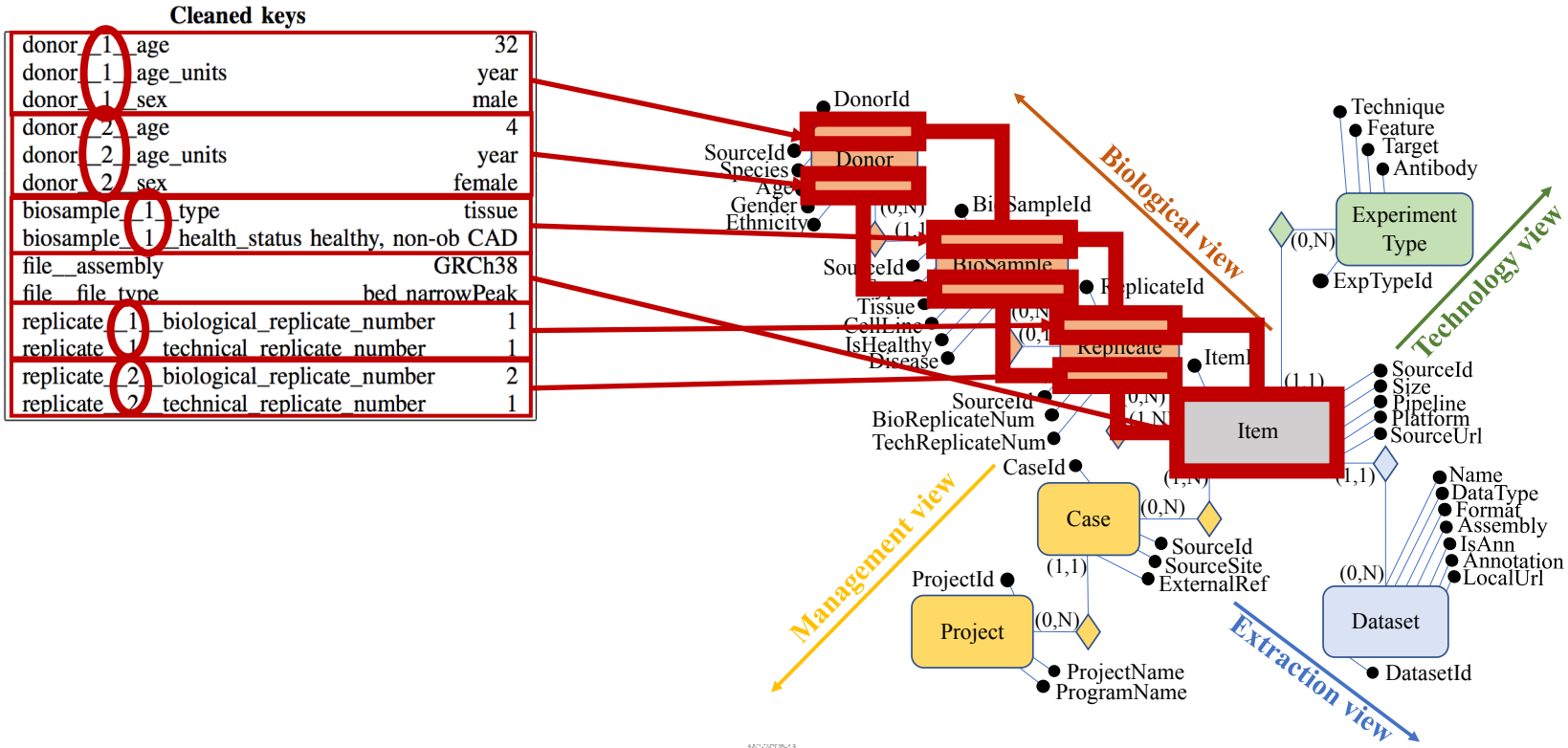
```
accession ENCSR635OSG
file__accession ENCFF429VMY
file__biological_replicates 1
file__biological_replicates 2
file__file_type bed narrowPeak

replicates__1__@id /replicates/4874c170-7124-4822-a058-4bb/
replicates__1__biological_replicate_number 1
replicates__1__library__biosample__donor__age 6
replicates__1__library__biosample__health_status healthy
replicates__2__@id /replicates/d42ff80d-67fd-45ee-9159-25a/
replicates__2__biological_replicate_number 2
replicates__2__library__biosample__donor__age 32
replicates__2__library__biosample__health_status
    healthy with non-obstructive coronary artery disease
```

Cleaner



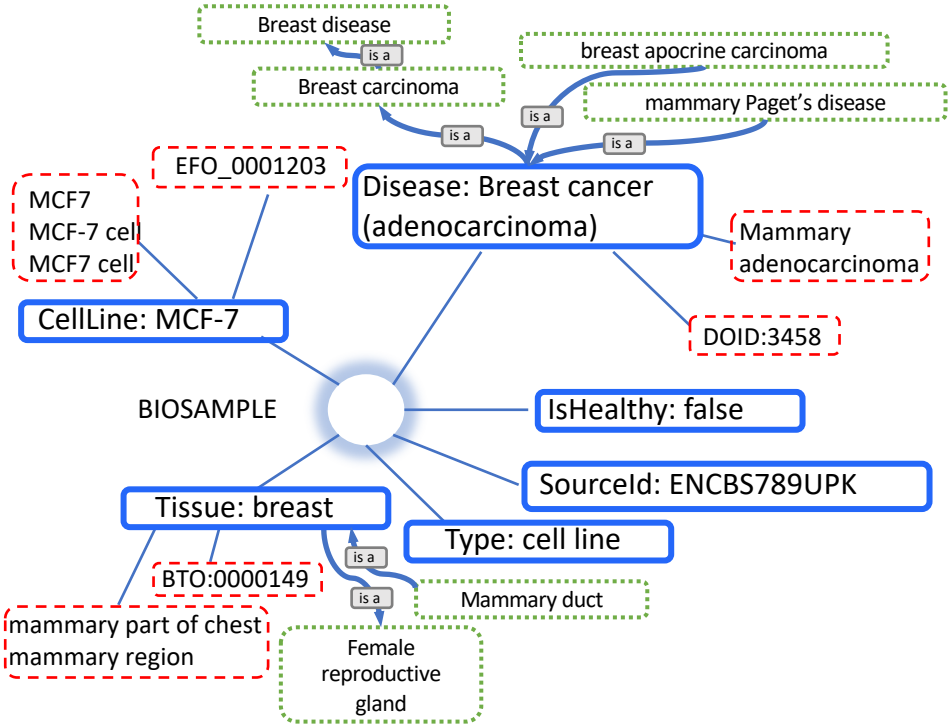
POLITECNICO
MILANO 1863

Mapper



POLITECNICO
MILANO 1863

Value Normalization and Enrichment



Normalization:

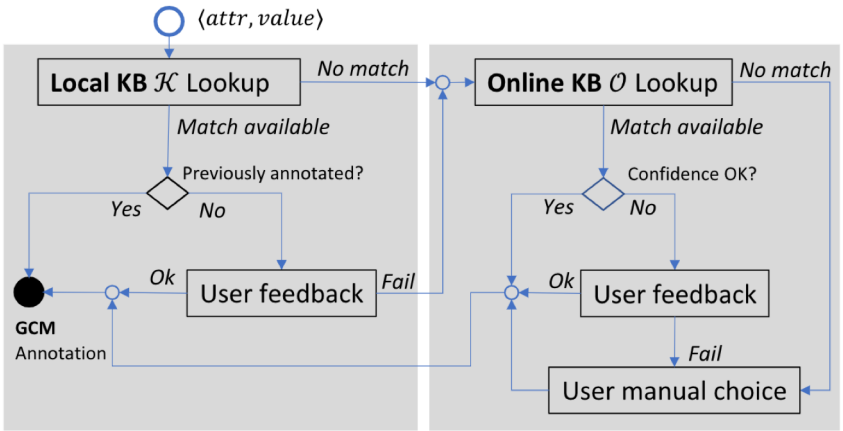
+ Ontology term id
+ Synonyms
+ External references to ontologies

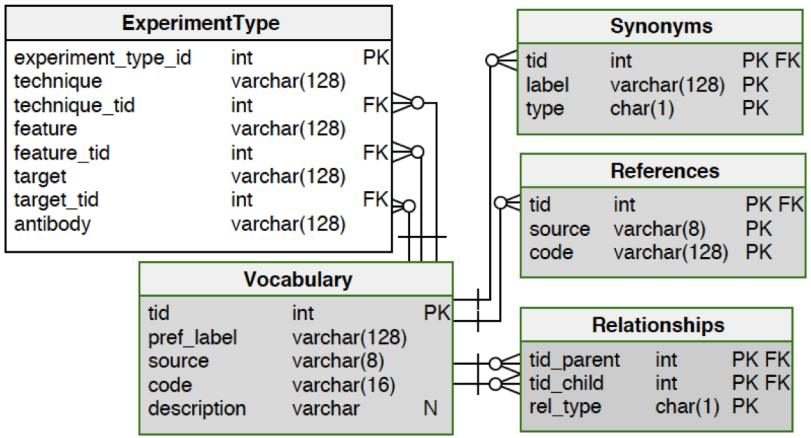
Enrichment:

+ Ancestors (hypernyms)
+ Descendants (hyponyms)

POLITECNICO
MILANO 1863

Value Normalization and Enrichment





Current GeCo repository

Stores experimental datasets and annotations collected from external databases:

ENCODE: more than 15,000 processed datasets for humans and mice, relevant to epigenomic research

Roadmap Epigenomics: about 2,000 human epigenomic datasets for stem cells and ex-vivo tissues

GDC (Genomic Data Commons): over 32,000 datasets including TCGA and TARGET programs, related to many aspects of cancer genomics:

- **TCGA** (The Cancer Genome Atlas): more than 11,000 processed datasets for ≈ 30 cancer types, including mutations, copy number variations, gene and miRNA expressions, methylations
- **TARGET** (Therapeutically Applicable Research To Generate Effective Treatments): ≈ 3,000 datasets applying a comprehensive genomic approach to determine molecular changes that drive childhood cancers

Source of data	Imported datasets	#samples	File size (MB)
ENCODE	GRCh38_ENCODE_BROAD	850	6,869
	GRCh38_ENCODE_NARROW	11,573	128,316
	HG19_ENCODE_BROAD	844	18,382
	HG19_ENCODE_NARROW	10,342	111,925
ROADMAP EPIGENOMICS	HG19_ROADMAP_EPIGENOMICS_BED	156	968
	HG19_ROADMAP_EPIGENOMICS_BROAD	979	24,332
	HG19_ROADMAP_EPIGENOMICS_DMR	66	3,060
	HG19_ROADMAP_EPIGENOMICS_GAPPED	979	6,875
	HG19_ROADMAP_EPIGENOMICS_NARROW	1,032	11,788
	HG19_ROADMAP_EPIGENOMICS_RNA_expression	399	2,453
TCGA	HG19_TCGA_cnv	22,632	797
	HG19_TCGA_dnamethylation	12,860	247,742
	HG19_TCGA_dnaseq	6,914	286
	HG19_TCGA_mirnaseq_isoform	9,909	4,207
	HG19_TCGA_mirnaseq_mirna	9,909	746
	HG19_TCGA_rnaseq_exon	3,675	47,668
	HG19_TCGA_rnaseq_gene	3,675	5,327
	HG19_TCGA_rnaseq_spljxn	3,675	44,377
	HG19_TCGA_rnaseqv2_exon	9,825	124,343
	HG19_TCGA_rnaseqv2_gene	9,825	21,862
	HG19_TCGA_rnaseqv2_isoform	9,825	53,082
	HG19_TCGA_rnaseqv2_spljxn	9,825	115,088
GDC - TCGA	GRCh38_TCGA_copy_number	22,374	686
	GRCh38_TCGA_copy_number_masked	22,375	337
	GRCh38_TCGA_gene_expression	11,091	56,542
	GRCh38_TCGA_methylation	12,218	1,348,516
	GRCh38_TCGA_miRNA_expression	10,947	1,502
	GRCh38_TCGA_miRNA_isoform_expression	10,999	5,004
	GRCh38_TCGA_somatic_mutation_masked	10,188	2,280
GENCODE	GRCh38_ANNOTATION_GENCODE	24	1,798
	HG19_ANNOTATION_GENCODE	20	1,324
REFSEQ	GRCh38_ANNOTATION_REFSEQ	31	740
	HG19_ANNOTATION_REFSEQ	30	275
Total	33 datasets	240,066	2,399,497

Try our system: http://geco.deib.polimi.it/repo-viewer/

DEMO TIME!