

# 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 6-8<sup>th</sup> 2019**

Organized by “GeCo” – Data-Driven Genomic Computing

# 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

Clear Disease Type IS Breast Invasive Carcinoma AND Primary Site IS Breast AND

Project Id IS TCGA-BRCA AND Data Category IS Simple Nucleotide Variation

Case UUID	Case ID	Project	Primary Site	Gender	Files
2779fa01-ac93-4e80-a997-3385f72172c3	<a href="#">TCGA-A8-A08S</a>	<a href="#">TCGA-BRCA</a>	Breast	Female	<a href="#">32</a>

Gene Expression Omnibus

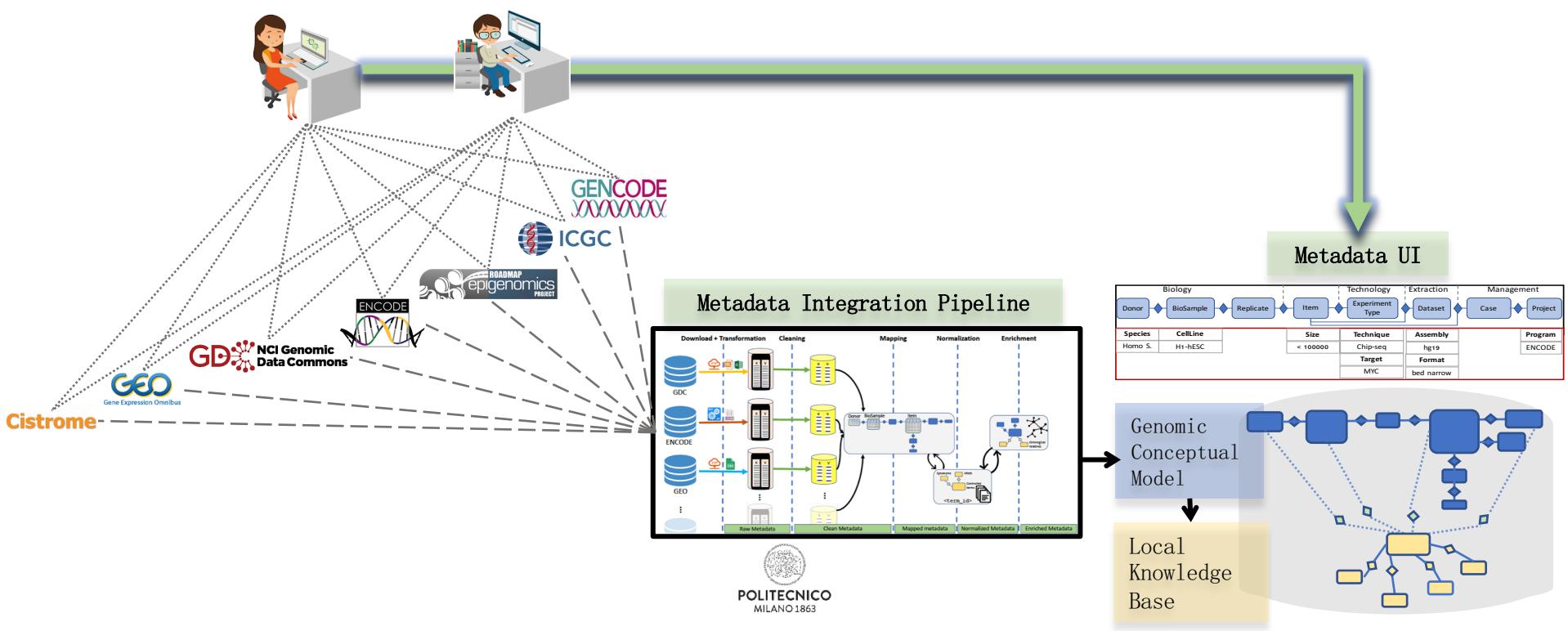
**Sample GSM1197482** Query DataSets for GSM1197482

Source name	T47D-MTVL
Organism	<a href="#">Homo Sapiens</a>
Characteristics	gender: female tissue: <u>breast cancer ductal carcinoma</u>

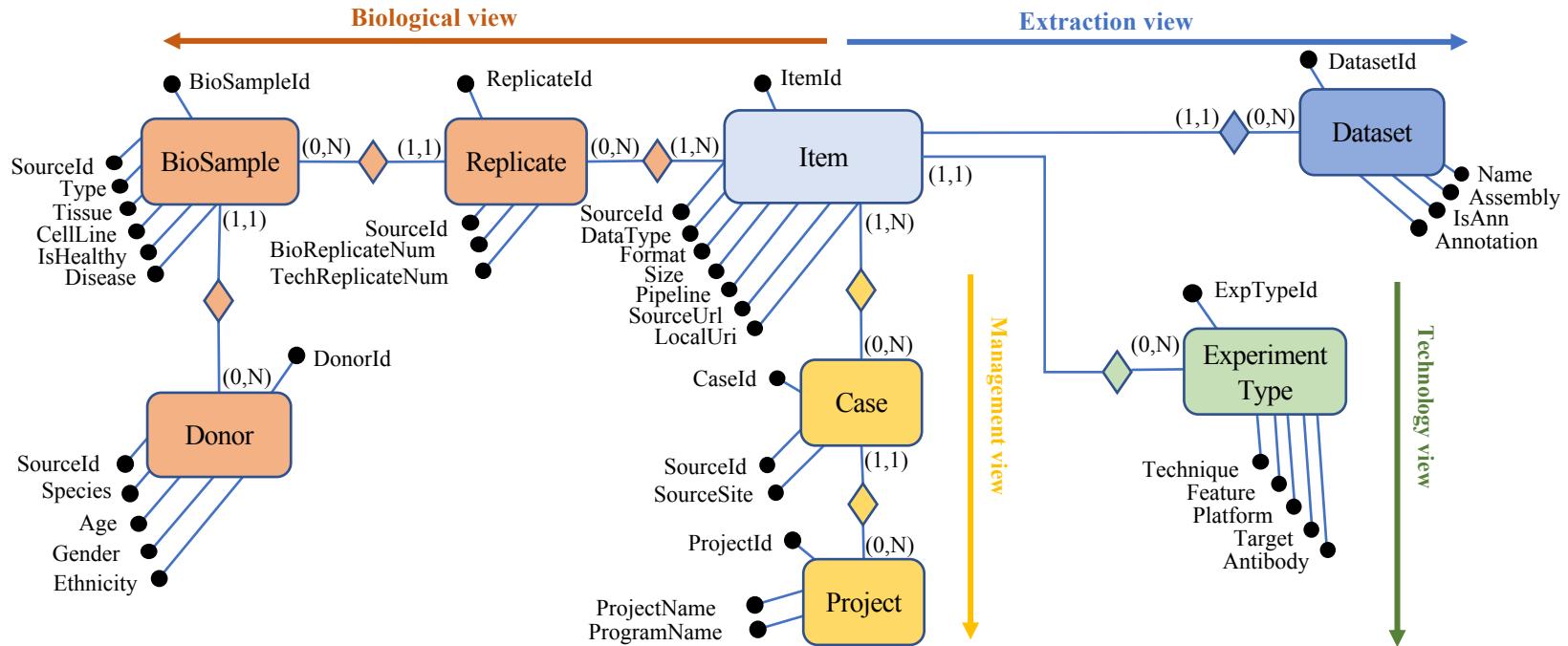
ENCODE

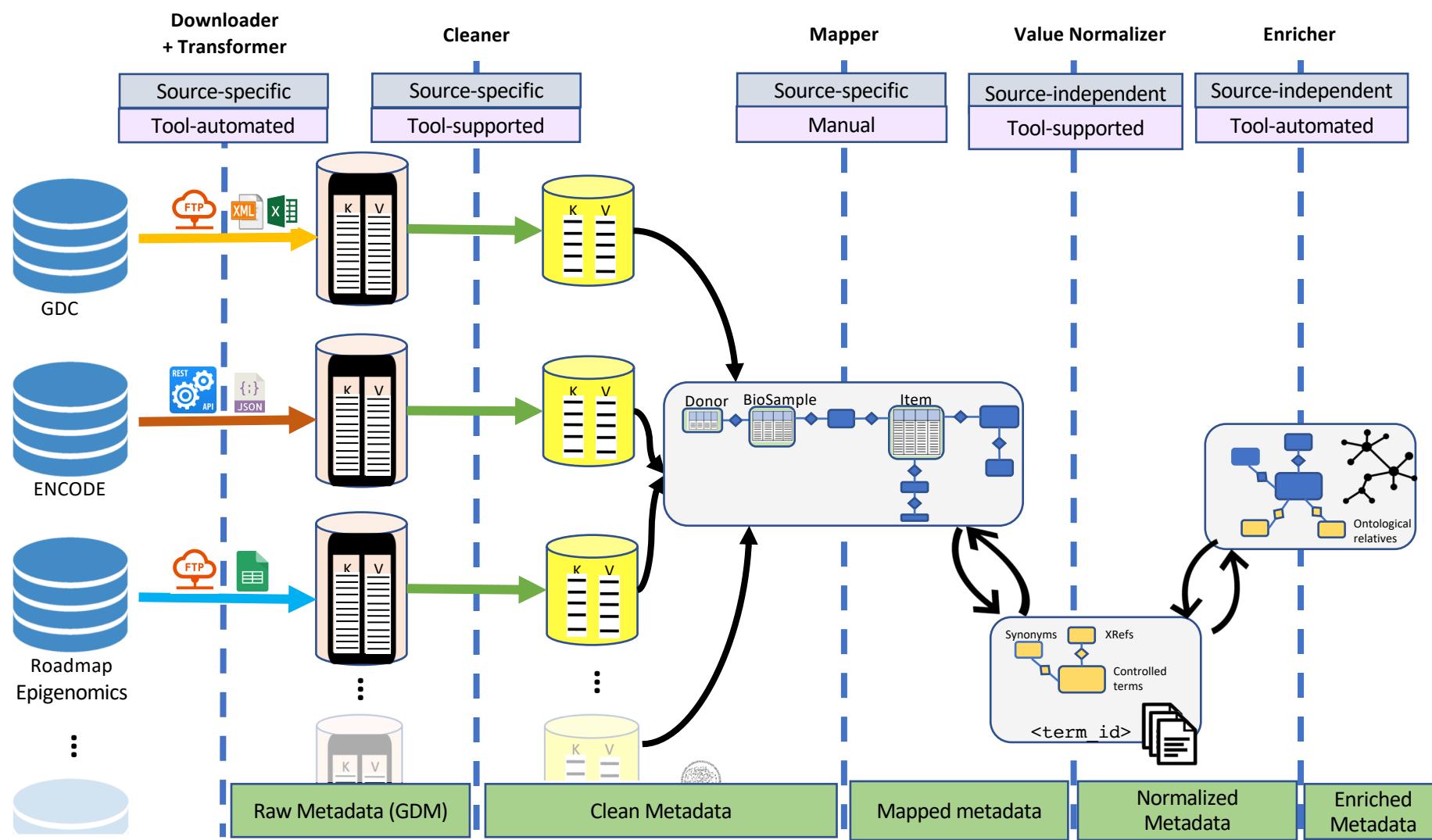
<b>Experiment summary for ENCSR000DMQ</b>	<b>Experiment summary for ENCSR000DOS</b>
Assay: ChIP-seq	Assay: ChIP-seq
Target: <a href="#">MYC</a>	Target: <a href="#">MYC</a>
Biosample: <i>Homo sapiens</i> MCF-7	Biosample: <i>Homo sapiens</i> MCF-10A
Biosample Type: cell line	Biosample Type: cell line
Description: <u>Mammary gland, adenocarcinoma</u>	Description: <u>Mammary gland, non-tumorigenic cell line</u>
Health status: <u>Breast cancer (adenocarcinoma)</u>	Health status: <u>Fibrocystic disease</u>

# Our approach



# The Genomic Conceptual Model (GCM)



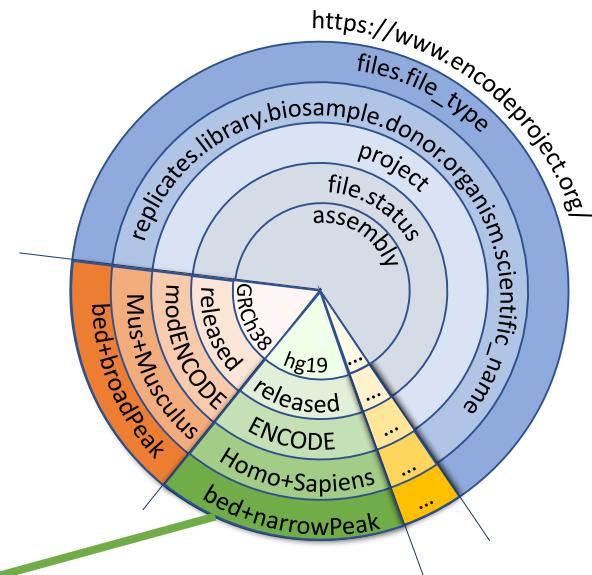


# Downloader

<https://www.encodeproject.org/metadata/?type=Experiment&{params}/metadata.tsv>

[https://www.encodeproject.org/files/{file\\_accession}/@@download/{file\\_accession}.bed.gz](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

```
chr1 237773 237868 1d-1 417 . -1 -1 -1 41,737
chr1 713878 743446 1d-2 1800 . -1 -1 -1 1800
chr1 762182 762415 1d-4 31 . -1 -1 -1 31,8939
chr1 762471 763071 1d-5 1800 . -1 -1 -1 1800
chr1 805184 805522 1d-7 331 . -1 -1 -1 33,8974
chr1 839695 848443 1d-8 893 . -1 -1 -1 893,3389
chr1 846748 846828 1d-10 15 . -1 -1 -1 15,4623
chr1 852854 852182 1d-11 18 . -1 -1 -1 18,46296
chr1 855911 856328 1d-13 25 . -1 -1 -1 25,52876
chr1 856433 856814 1d-14 367 . -1 -1 -1 36,6562
chr1 860869 860869 1d-15 29 . -1 -1 -1 29,68223
chr1 860869 864408 1d-16 59 . -1 -1 -1 5,86284
chr1 871246 875528 1d-17 55 . -1 -1 -1 5,46648
chr1 877828 877536 1d-19 82 . -1 -1 -1 82,23582
chr1 880073 880196 1d-20 17 . -1 -1 -1 1,66535
chr1 885798 886321 1d-21 183 . -1 -1 -1 183,42389
chr1 895798 896532 1d-22 182 . -1 -1 -1 182,2851
chr1 896736 896932 1d-23 24 . -1 -1 -1 24,41455
chr1 913554 913662 1d-25 48 . -1 -1 -1 48,58648
chr1 913554 918664 1d-25 86 . -1 -1 -1 86,58648
chr1 933554 933662 1d-26 16 . -1 -1 -1 15,55863
chr1 944513 948652 1d-28 28 . -1 -1 -1 28,78274
chr1 947723 948692 1d-29 21 . -1 -1 -1 21,18759
chr1 949812 949909 1d-31 13 . -1 -1 -1 13,54889
chr1 954477 956343 1d-32 1800 . -1 -1 -1 1800,9,9421
chr1 956688 957308 1d-33 19 . -1 -1 -1 19,19898
chr1 966688 966916 1d-35 22 . -1 -1 -1 22,16895
chr1 968688 969008 1d-36 1980 . -1 -1 -1 1980,1988
chr1 989818 979236 1d-37 88 . -1 -1 -1 88,7929
chr1 974613 974851 1d-38 49 . -1 -1 -1 49,485527
chr1 975682 975681 1d-39 14 . -1 -1 -1 14,21744
chr1 975682 975681 1d-40 14 . -1 -1 -1 14,43383
chr1 975864 976849 1d-41 776 . -1 -1 -1 77,5736
chr1 975864 976849 1d-42 14 . -1 -1 -1 14,14282
chr1 994646 994773 1d-43 20 . -1 -1 -1 20,1,96167
chr1 994840 995346 1d-45 489 . -1 -1 -1 48,48596
chr1 998252 998439 1d-46 22 . -1 -1 -1 22,2,21413
chr1 999237 999434 1d-47 58 . -1 -1 -1 58,4,99733
chr1 1003958 1005179 1d-48 48 . -1 -1 -1 48,10298
chr1 1003958 1005179 1d-49 88 . -1 -1 -1 88,7526
chr1 1014997 1014965 1d-50 141 . -1 -1 -1 141,1111
chr1 1015185 1025419 1d-52 1000 . -1 -1 -1 1000,10,93556
chr1 1015185 1025419 1d-52 1000 . -1 -1 -1 1000,9,93556
chr1 1075717 1075832 1d-53 194 . -1 -1 -1 194,9,93556
chr1 1079691 1079709 1d-54 21 . -1 -1 -1 21,2,89496
chr1 1092684 1093206 1d-55 37 . -1 -1 -1 37,1,91765
chr1 1092684 1093206 1d-56 14 . -1 -1 -1 14,1,91765
chr1 1093491 1093978 1d-59 74 . -1 -1 -1 74,41889
```

```
accession ENCSCR6350SG
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

## Transformed keys

replicates_1_library_biosample_donor_age	32
replicates_1_library_biosample_donor_age_units	year
replicates_1_library_biosample_donor_sex	male
replicates_2_library_biosample_donor_age	4
replicates_2_library_biosample_donor_age_units	year
replicates_2_library_biosample_donor_sex	female
replicates_1_library_biosample_sex	male
replicates_1_library_biosample_type	tissue
replicates_1_library_biosample_health_status	healthy, non-ob CAD
file_biological_replicate	1
file_technical_replicate	1_1
file_assembly	GRCh38
file_file_type	bed narrowPeak
replicates_1_biological_replicates	
replicates_1_technical_replicates	
replicates_2_biological_replicates	
replicates_2_technical_replicates	
assembly	
assembly	

replicates(\_[0-9]\_)library\_biosample\_(donor)\_(\_age|sex)(.\* ) => \$2\$1\$3\$4

## RuleBase

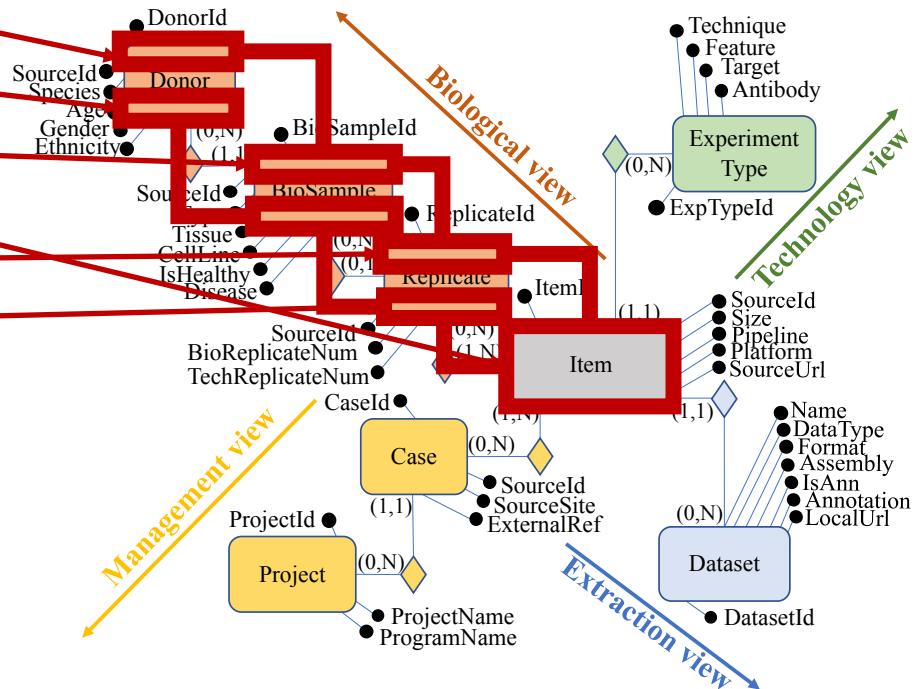
- (2) replicates\_[0-9]\_library\_biosample\_sex.\* => DELETE
- (3) replicates(\_[0-9]\_)library\_(biosample)\_(\_biosample\_)?(.\* )  
=> \$2\$1\$4
- (4) file\_(biologicaltechnical)\_replicates => DELETE
- (5) (file\_)(file\_)?(.\* ) => \$1\$3
- (6) (replicate)s(\_[0-9]\_)(.\* ) => \$1\$2\$3
- (7) assembly => DELETE

## Cleaned keys

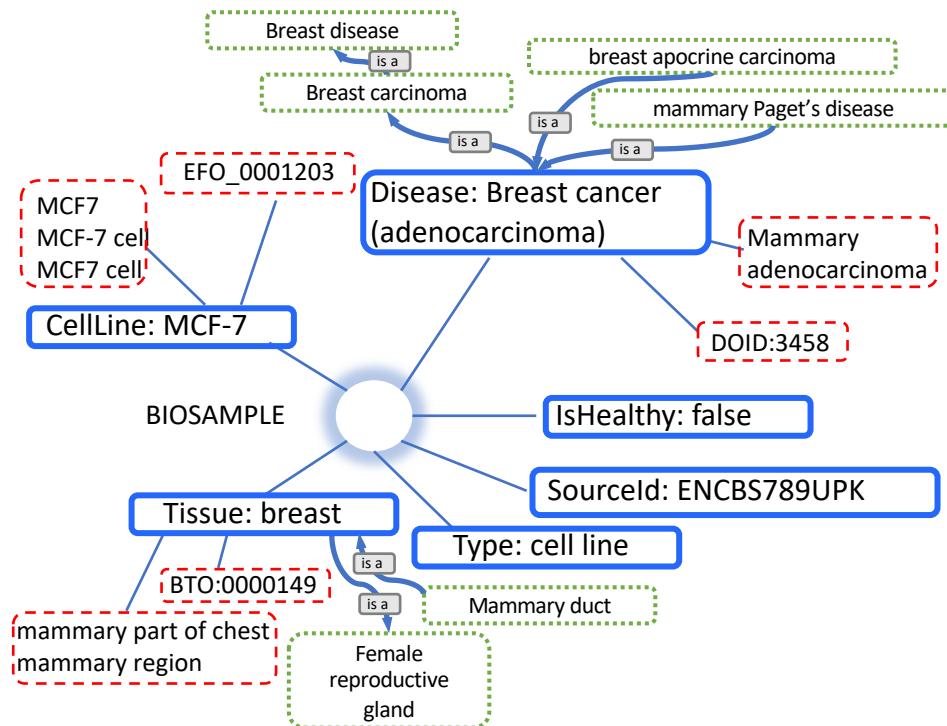
donor_1_age	32
donor_1_age_units	year
donor_1_sex	male
donor_2_age	4
donor_2_age_units	year
donor_2_sex	female
biosample_1_type	tissue
biosample_1_health_status	healthy, non-ob CAD
file_assembly	GRCh38
file_file_type	bed narrowPeak
replicate_1_biological_replicates	1
replicate_1_technical_replicates	1
replicate_2_biological_replicates	2
replicate_2_technical_replicates	1

# Mapper

Cleaned keys	
donor_1_age	32
donor_1_age_units	year
donor_1_sex	male
donor_2_age	4
donor_2_age_units	year
donor_2_sex	female
biosample_1_type	tissue
biosample_1_health_status	healthy, non-ob CAD
file_assembly	GRCh38
file_file_type	bed narrowPeak
replicate_1_biological_replicate_number	1
replicate_1_technical_replicate_number	1
replicate_2_biological_replicate_number	2
replicate_2_technical_replicate_number	1



# Value Normalization and Enrichment



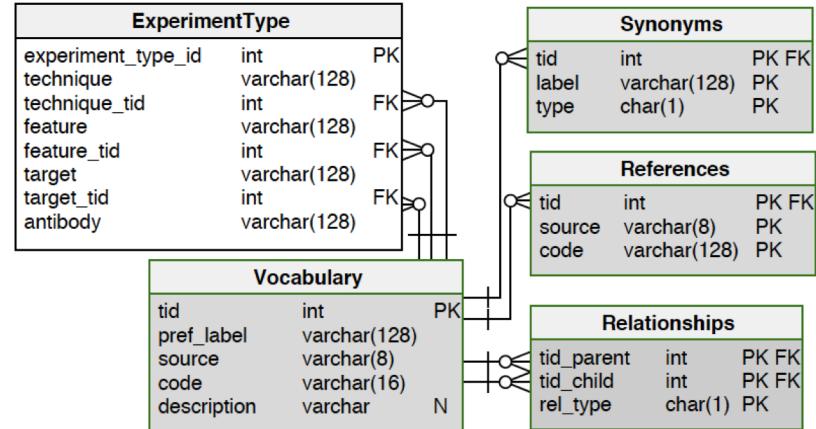
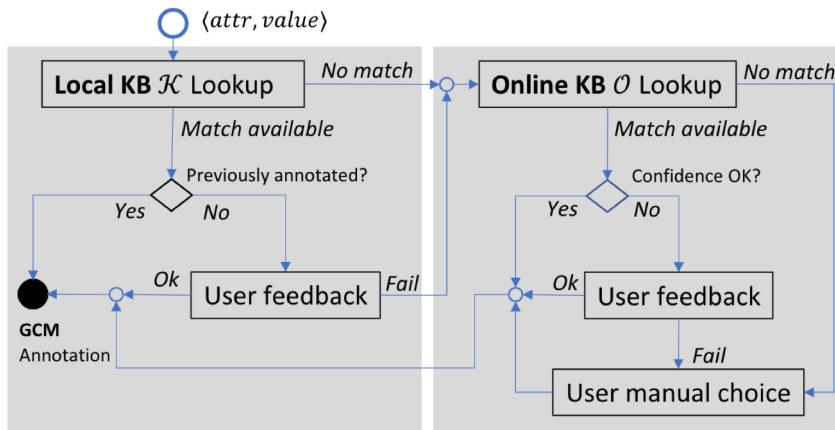
Normalization:

- + Ontology term id
- + Synonyms
- + External references to ontologies

Enrichment:

- + Ancestors (hypernyms)
- + Descendants (hyponyms)

# 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  $\approx$  30 cancer types, including mutations, copy number variations, gene and miRNA expressions, methylations
- **TARGET** (Therapeutically Applicable Research To Generate Effective Treatments):  $\approx$  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_mnaseq_exon	3,675	47,668
	HG19_TCGA_mnaseq_gene	3,675	5,327
	HG19_TCGA_mnaseq_spljxn	3,675	44,377
	HG19_TCGA_mnaseqv2_exon	9,825	124,343
	HG19_TCGA_mnaseqv2_gene	9,825	21,862
	HG19_TCGA_mnaseqv2_isoform	9,825	53,082
	HG19_TCGA_mnaseqv2_spljxn	9,825	115,088
	GRCh38_TCGA_copy_number	22,374	686
	GRCh38_TCGA_copy_number_masked	22,375	337
GDC - TCGA	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!

