

Managing and leveraging knowledge catalogs with TKCat

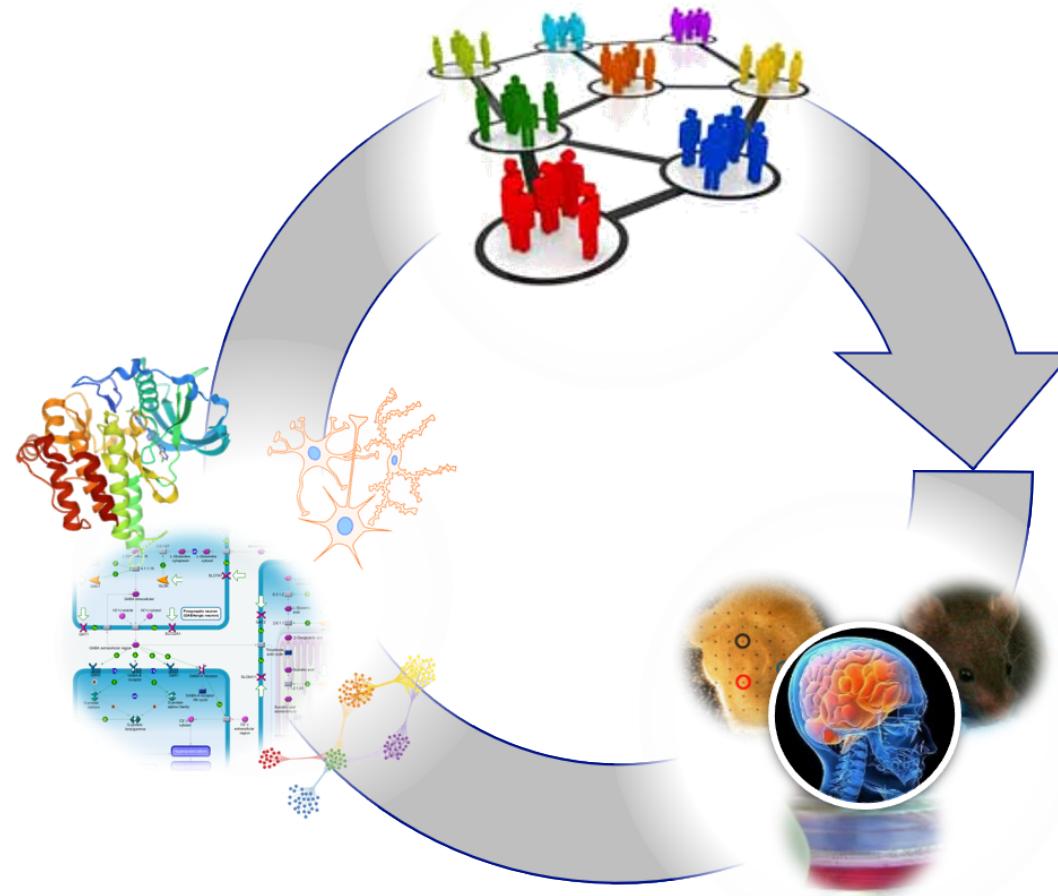
Patrice Godard | useR!2022 | 23 Jun 2022



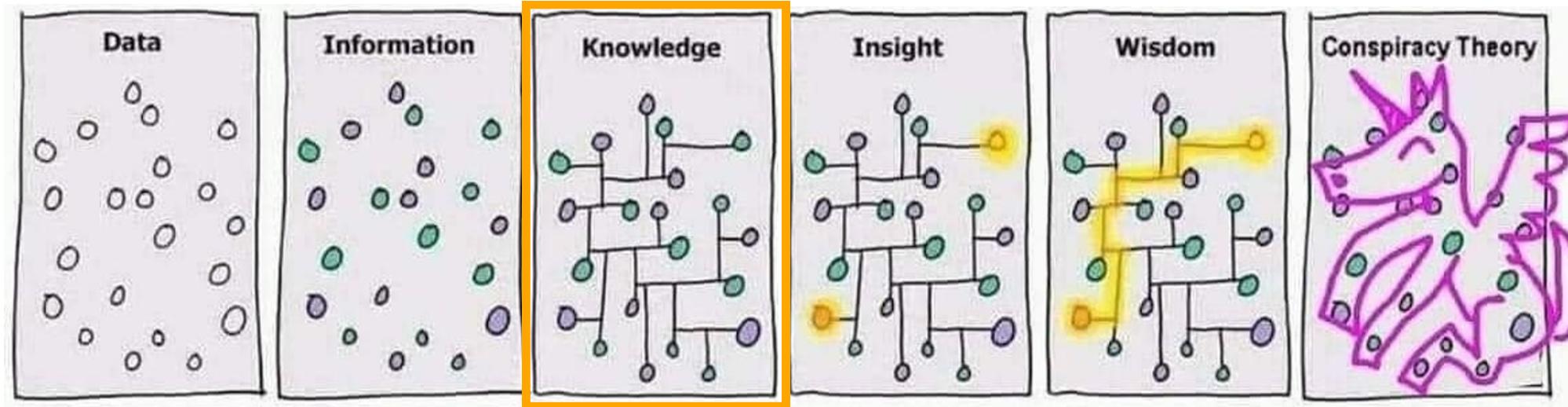
Inspired by **patients**.
Driven by **science**.



Translational Bioinformatics at UCB

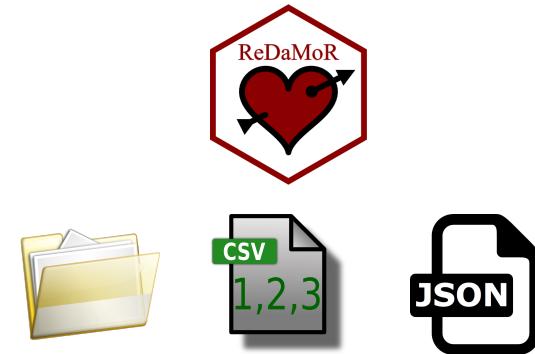


From data to wisdom



Expected features of the knowledge to manage

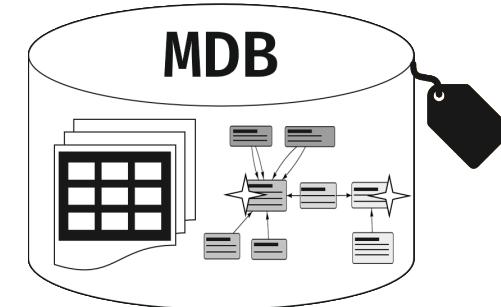
- Diverse concepts
- Connected concepts
- Tabular data
- To use in R and beyond
- Non monolytic knowledge: management of independent pieces
- Ready for integration on the basis of shared concepts
- Diverse implementation of shared concepts
- Potentially billions of records
- Tables to be used entirely or by subset
- No frequent updates but to be versioned
- Potential restriction of use



MDB: a Modeled Database for each knowledge resource

Features

- **Data:** tables and matrices
- **Data model:** Formal description of the data
- **Description:** general information about the data
- **Collections:** tables referring to key concepts
- **Implementation:** in memory, in files, in ClickHouse



TKCat

- **Tailored Knowledge Catalog:** a package for managing and using MDBs



Drafting a data model in R with ReDaMoR

```
library(readr)
hpo_data_dir <- system.file("examples/HPO-subset", package="ReDaMoR")
HPO_hp <- read_tsv(file.path(hpo_data_dir, "HPO_hp.txt"))
HPO_diseases <- read_tsv(file.path(hpo_data_dir, "HPO_diseases.txt"))
HPO_diseaseHP <- read_tsv(file.path(hpo_data_dir, "HPO_diseaseHP.txt"))
```

```
library(ReDaMoR)
hpo_model <- df_to_model(HPO_hp, HPO_diseases, HPO_diseaseHP)
plot(hpo_model)
```

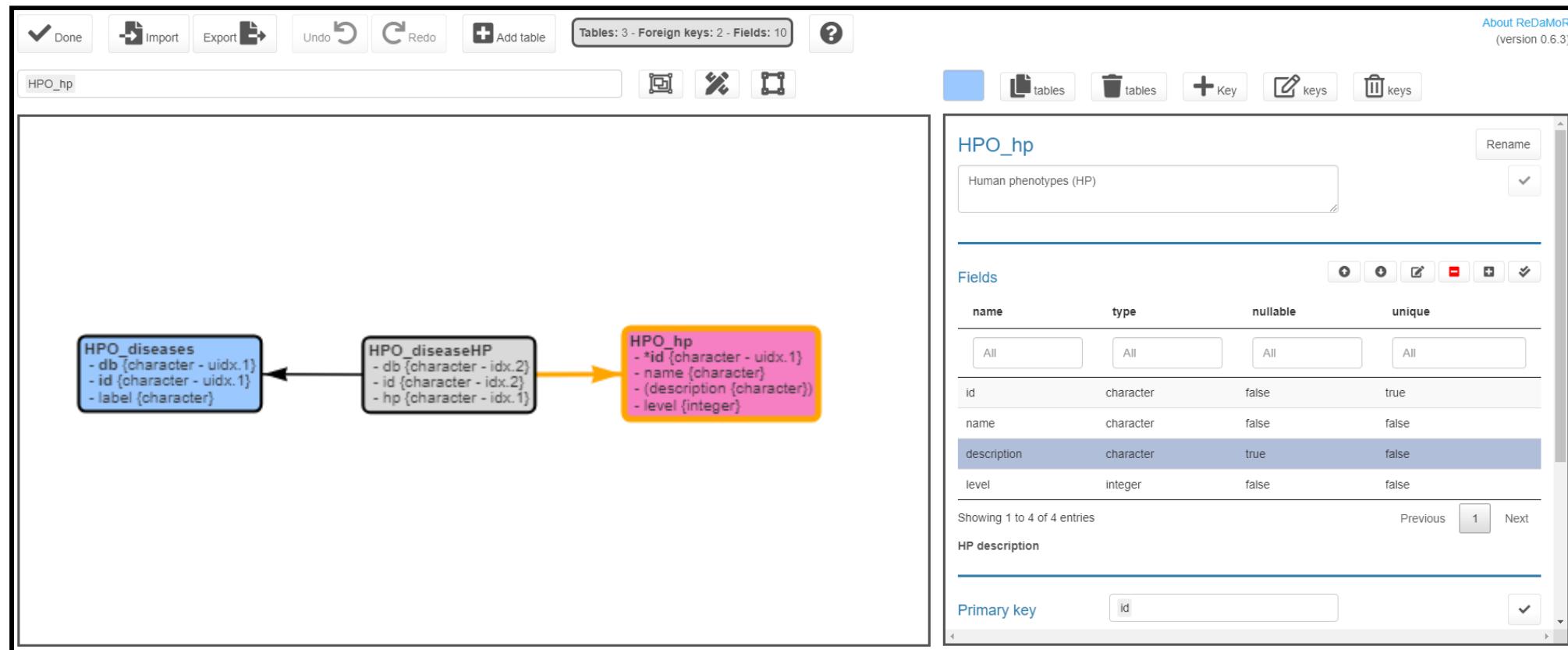
HPO_hp
- (id {character})
- (name {character})
- (description {character})
- (level {numeric})

HPO_diseases
- (db {character})
- (id {numeric})
- (label {character})

HPO_diseaseHP
- (db {character})
- (id {numeric})
- (hp {character})

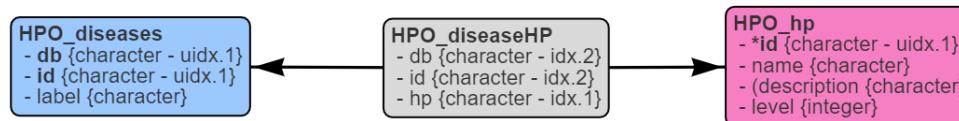
Creating a data model in R with ReDaMoR

```
hpo_model <- model_relational_data(hpo_model)
```



Confronting data to the model

```
confront_data(hpo_model, data=list(
  "HPO_hp"=HPO_hp,
  "HPO_diseaseHP"=HPO_diseaseHP,
  "HPO_diseases"=HPO_diseases
))
```



```
HP0_hp <- mutate(HP0_hp, level=as.integer(level))
HP0_diseases <- mutate(HP0_diseases, id=as.character(id))
HP0_diseaseHP <- mutate(HP0_diseaseHP, id=as.character(id))
confront_data(hpo_model, data=list(
  "HPO_hp"=HPO_hp,
  "HPO_diseaseHP"=HPO_diseaseHP,
  "HPO_diseases"=HPO_diseases
))
```

Confrontation with original data

FAILURE

Check configuration

- Optional checks: unique, not nullable, foreign keys
- Maximum number of records: Inf

HPO_hp

FAILURE

Field issues or warnings

- description: SUCCESS Missing values 117/500 = 23%
- level: FAILURE Unexpected "numeric"

HPO_diseases

FAILURE

Field issues or warnings

- id: FAILURE Unexpected "numeric"

HPO_diseaseHP

FAILURE

Field issues or warnings

- id: FAILURE Unexpected "numeric"

Confrontation with corrected data

SUCCESS

Check configuration

- Optional checks: unique, not nullable, foreign keys
- Maximum number of records: Inf

HPO_hp

SUCCESS

Field issues or warnings

- description: SUCCESS Missing values 117/500 = 23%

Creating and using an MDB with TKCat

MDB creation

```
library(TKCat)
hpo <- memoMDB(
  dataTables=list(
    "HPO_hp"=HPO_hp,
    "HPO_diseases"=HPO_diseases,
    "HPO_diseaseHP"=HPO_diseaseHP
  ),
  dataModel=hpo_model,
  dbInfo=list(
    name="miniHPO",
    title="Very small extract of the human phenome",
    description="For demonstrating ReDaMoR and TKCat",
    url="https://hpo.jax.org/app/",
    version="0.1",
    maintainer="Patrice Godard <patrice.godard@inserm.fr>"
  )
)
```

Explore and retrieve information

```
db_info(hpo)
data_model(hpo)
```

```
hpo %>% select(HPO_diseases, HPO_diseaseHP)
hpo %>% pull(HPO_diseases) %>%
  head(3)

## # A tibble: 3 × 3
##   db      id    label
##   <chr>  <chr> <chr>
## 1 DECIPHER 15    NF1-microdeletion syndrome
## 2 DECIPHER 45    Xq28 (MECP2) duplication
## 3 DECIPHER 65    ATR-16 syndrome
```

Leverage the MDB data model: filter

```
dims(hpo) %>% select(name, nrow)
```

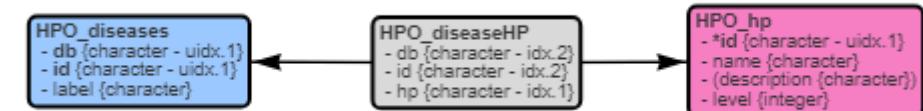
```
## # A tibble: 3 × 2
##   name      nrow
##   <chr>     <int>
## 1 HPO_hp    500
## 2 HPO_diseases 1903
## 3 HPO_diseaseHP 2594
```

```
fhp0 <- hpo %>% filter(HPO_hp=stringr::str_detect(description, "eye"))
```

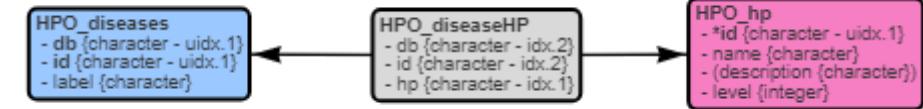
```
fhp0 %>% dims() %>% select(name, nrow)
```

```
## # A tibble: 3 × 2
##   name      nrow
##   <chr>     <int>
## 1 HPO_hp    21
## 2 HPO_diseaseHP 109
## 3 HPO_diseases 99
```

```
data_model(hpo) %>% plot()
```



```
data_model(fhp0) %>% plot()
```

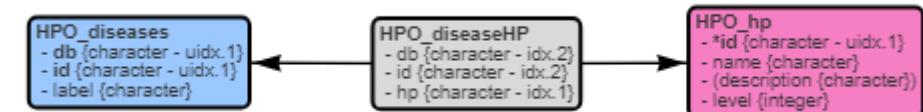


Leverage the MDB data model: join

```
dims(fhpo) %>% select(name, nrow)

## # A tibble: 3 × 2
##   name      nrow
##   <chr>     <int>
## 1 HPO_hp     21
## 2 HPO_diseaseHP 109
## 3 HPO_diseases    99
```

```
data_model(fhpo) %>% plot()
```

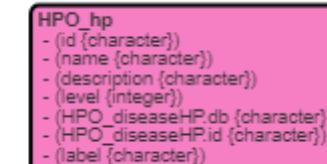


```
jhpo <- fhpo %>% join_mdb_tables(c("HPO_hp", "HPO_diseaseHP", "HPO_diseases"))
```

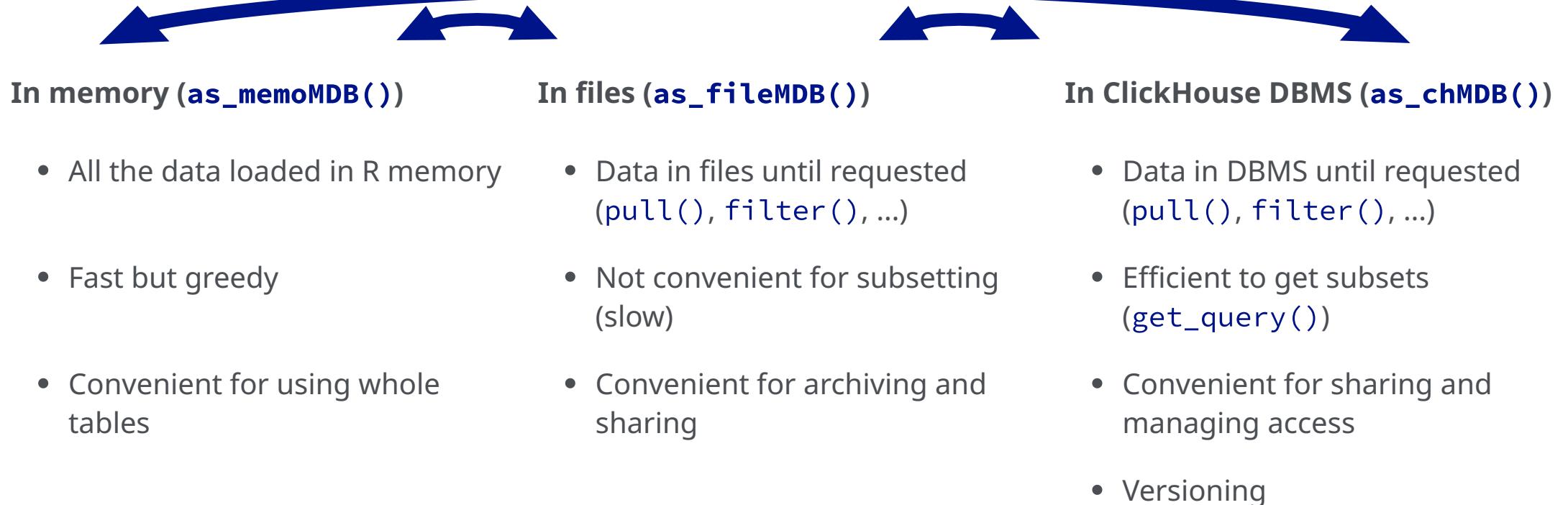
```
data_model(jhpo) %>% plot()
```

```
jhpo %>% dims() %>% select(name, nrow)
```

```
## # A tibble: 1 × 2
##   name      nrow
##   <chr>     <int>
## 1 HPO_hp     109
```



MDB implementations



TKCat: a data warehouse management system

```
k <- chTKCat()
  host="local"
  user="def"
  password=
)
explore_MDBs
```

chTKCat : UCB - TBN

ClinVar_entrezNames

Collection members

collection	id	table	field	static	value	type
BE	1	ClinVar_entrezNames	be	true	Gene	
BE	1	ClinVar_entrezNames	identifier	false	entrez	
BE	1	ClinVar_entrezNames	organism	true	Homo sapiens	Scientific name
Condition	1	ClinVar_traitCref	condition	true	Disease	
Condition	1	ClinVar_traitCref	identifier	false	id	
Condition	1	ClinVar_traitCref	source	false	db	

Showing 1 to 7 of 10 entries

ClinVar_entrezNames

- Number of records: 36,024 (showing 100 records)

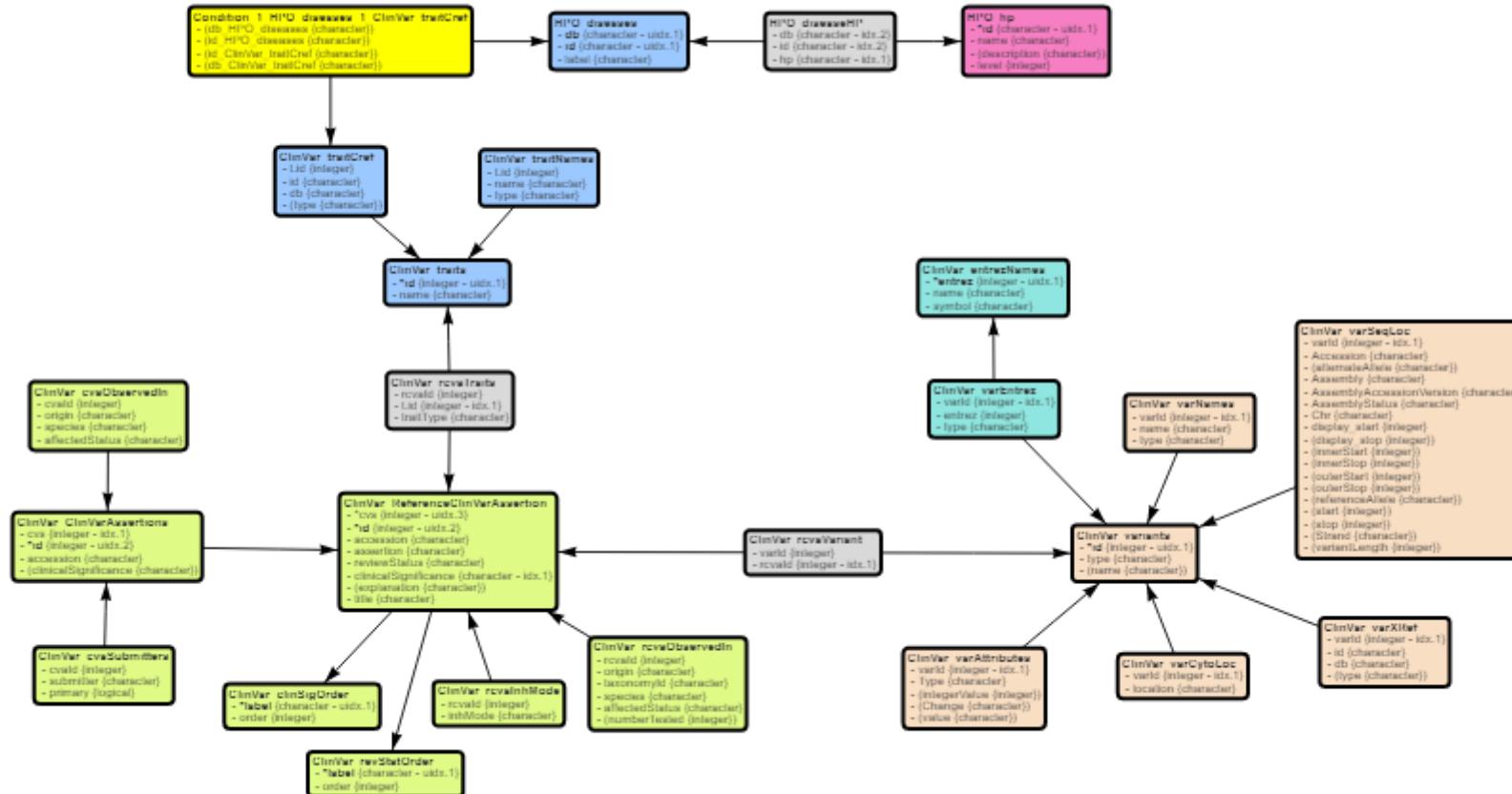
entrez	name	symbol
1	alpha-1-B glycoprotein	A1BG
2	alpha-2-macroglobulin	A2M
9	N-acetyltransferase 1	NAT1
10	N-acetyltransferase 2	NAT2
12	serpin family A member 3	SERPIN
13	arylacetamide deacetylase	AADAC

Showing 1 to 7 of 100 entries

Request table for download (Preparing the data may take time)

Previous 1 2 3 Next

Merging MDBs with collections



Supported data types

- character, numeric, integer, logic, Date, POSIXct (time)

- base64 (file)

MetaBase_Imagemaps

id	pngB64	name
5	file	Apoptosis and survival_Regulation of apoptosis by mitochondrial proteins
138	file	Regulation of lipid metabolism_Regulation of acetyl-CoA carboxylase 1 activity
140	file	Regulation of lipid metabolism_Regulation of acetyl-CoA carboxylase 2 activity in muscle
143	file	Regulation of lipid metabolism_Regulation of fatty acid synthase activity in hepatocytes
369	file	Transcription_Mechanism of activation of the transcription of Retinoid-target genes

- matrix and sparse matrix

Normalized_single_nuc

Number of records : 6,773,962,864 (showing 100 records)

gene	cell	value
ENSMUSG00000051951	P1_P1_SAMPLE_AAACCCATCAGAGCAG-1	1.6505107700341
ENSMUSG00000089699	P1_P1_SAMPLE_AAAGAACAGGTCCCT	0
ENSMUSG00000102331	P1_P1_SAMPLE_AAAGAACAGGTCCCT	0



<https://patzaw.github.io/TKCat-useR2022/TKCat-useR2022-Patrice-Godard.html>