HelmholtzZentrum münchen
German Research Center for Environmental Health

ICB Institute of Computational Biology

Neo4j - a graph DB

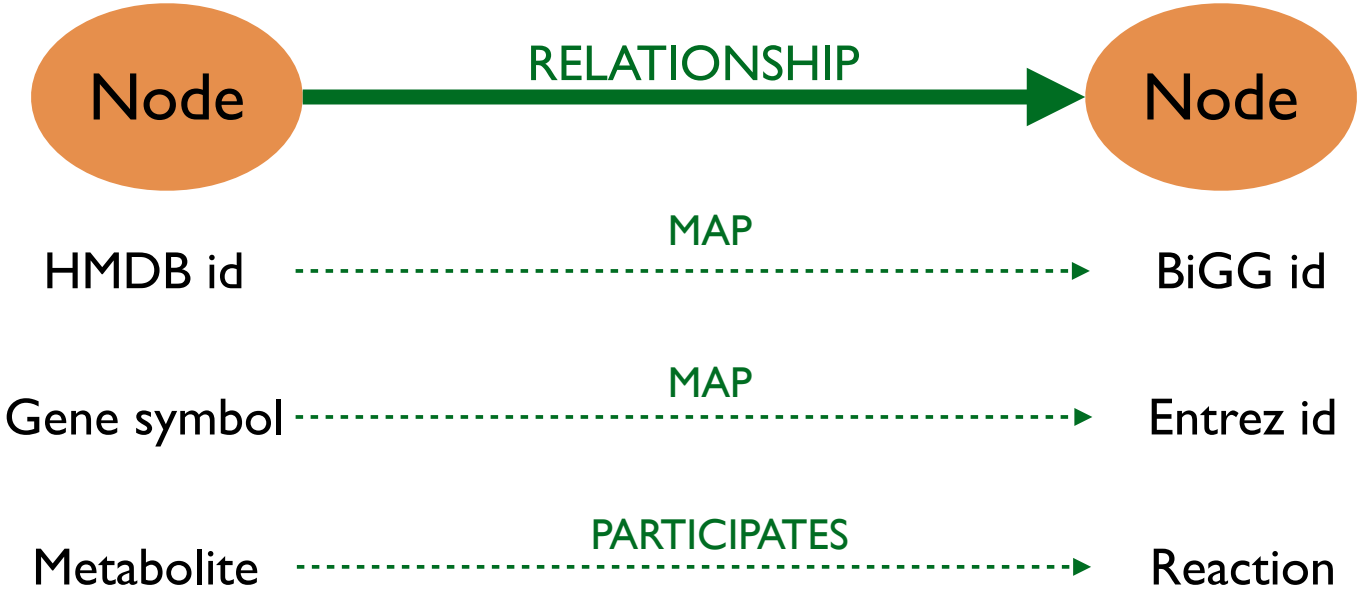
Maria Wörheide July 27, 2018

Overview

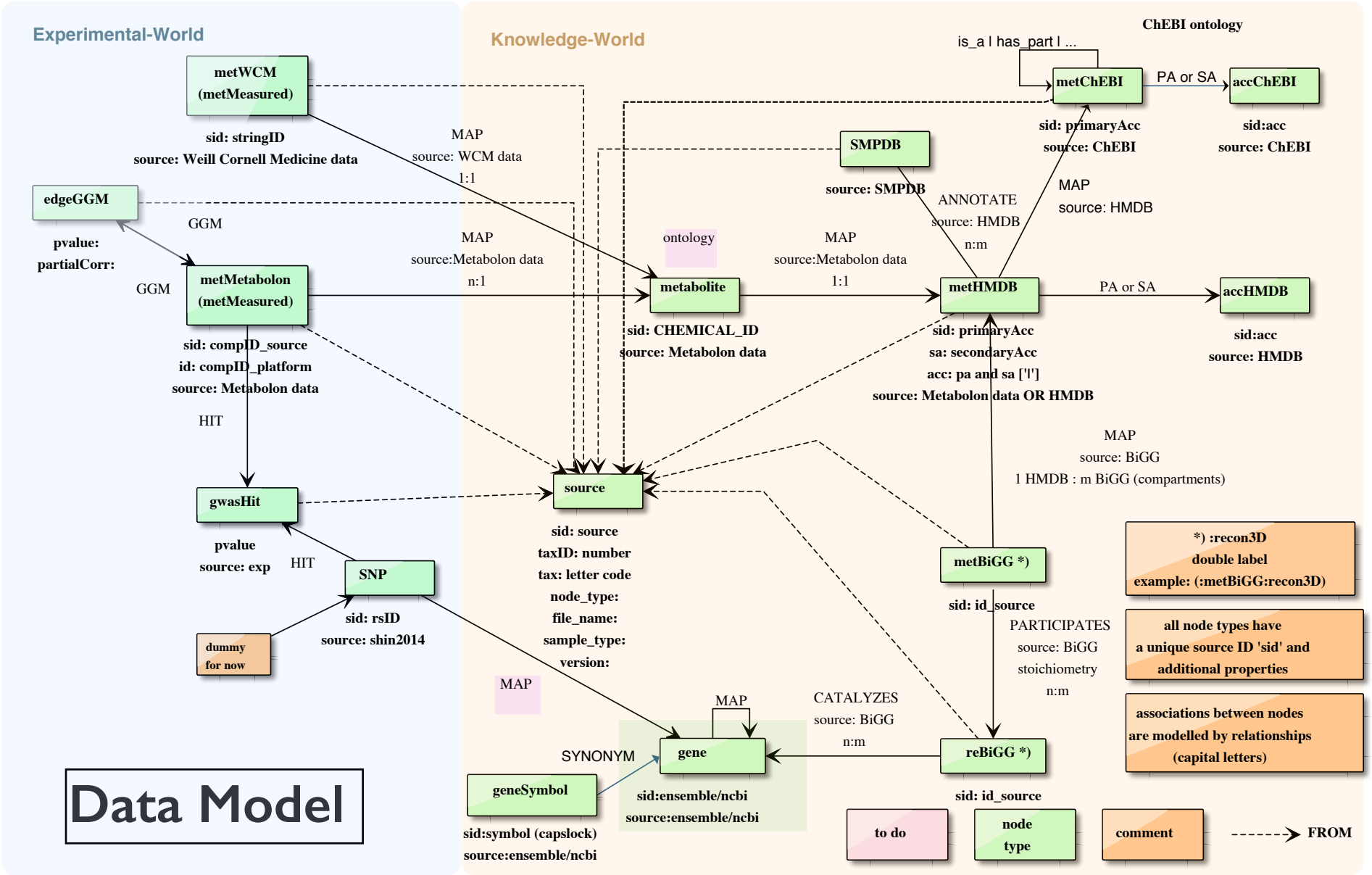
- Introduction to Neo4j and Cypher
- Data model
- Git repository
- Neo4j Servers
- Code

neo4j - a native graph DB

“ Today's world is no longer driven by data – it's driven by the connections between them[1]. “



[1]https://neo4j.com/product/


Experimental-World
Knowledge-World
ChEBI ontology
is_a I has_part I ...
metWCM
(metMeasured)
sid: stringID
source: Weill Cornell Medicine data
MAP
source: WCM data
1:1
edgeGGM
pvalue:
partialCorr:
GGM
GGM
metMetabolon
(metMeasured)
sid: compID_source
id: compID_platform
source: Metabolon data
MAP
source:Metabolon data
n:1
ontology
metabolite
sid: CHEMICAL_ID
source: Metabolon data
MAP
source:Metabolon data
1:1
SMPDB
source: SMPDB
ANNOTATE
source: HMDB
n:m
MAP
source: HMDB
metChEBI
sid: primaryAcc
source: ChEBI
PA or SA
accChEBI
sid:acc
source: ChEBI
metHMDB
sid: primaryAcc
sa: secondaryAcc
acc: pa and sa ['|']
source: Metabolon data OR HMDB
PA or SA
accHMDB
sid:acc
source: HMDB
HIT
gwasHit
pvalue
source: exp
HIT
SNP
sid: rsID
source: shin2014
dummy
for now
source
sid: source
taxID: number
tax: letter code
node_type:
file_name:
sample_type:
version:
MAP
source: BiGG
1 HMDB : m BiGG (compartments)
metBiGG *)
sid: id_source
PARTICIPATES
source: BiGG
stoichiometry
n:m
*) :recon3D
double label
example: (:metBiGG:recon3D)
all node types have
a unique source ID 'sid' and
additional properties
associations between nodes
are modelled by relationships
(capital letters)
MAP
MAP
CATALYZES
source: BiGG
n:m
SYNONYM
gene
sid:ensemble/ncbi
source:ensemble/ncbi
geneSymbol
sid:symbol (capslock)
source:ensemble/ncbi
reBiGG *)
sid: id_source
to do
node
type
comment
FROM
Data Model

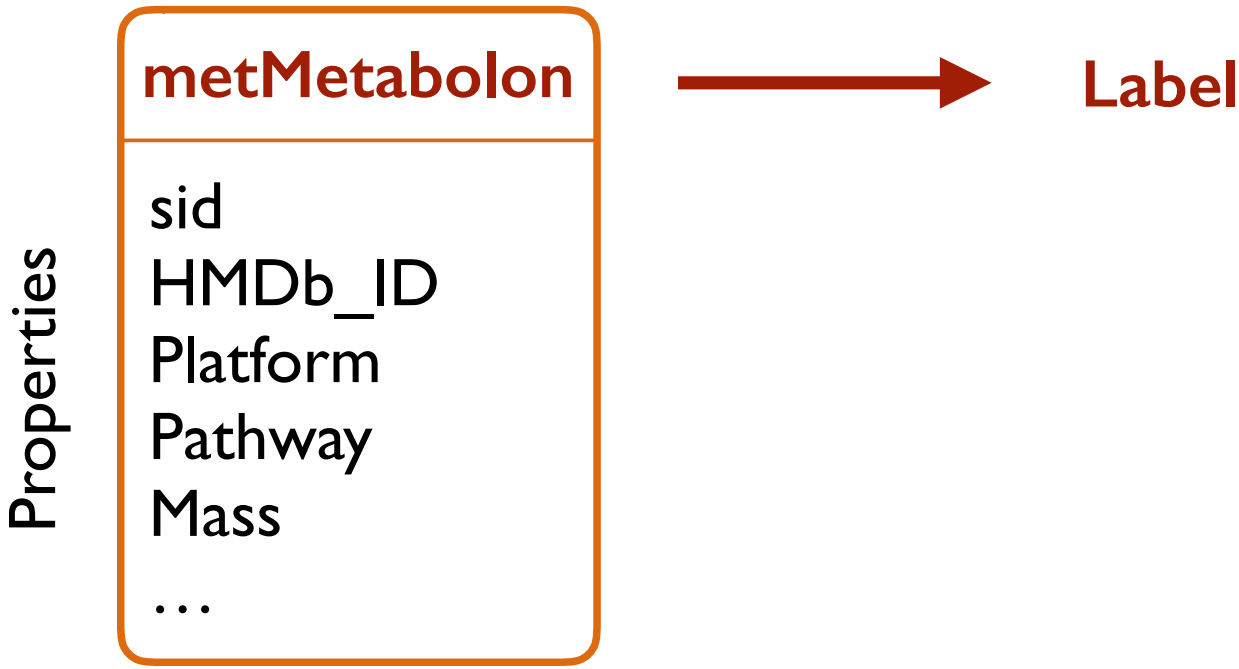
Neo4j - Nodes

Properties

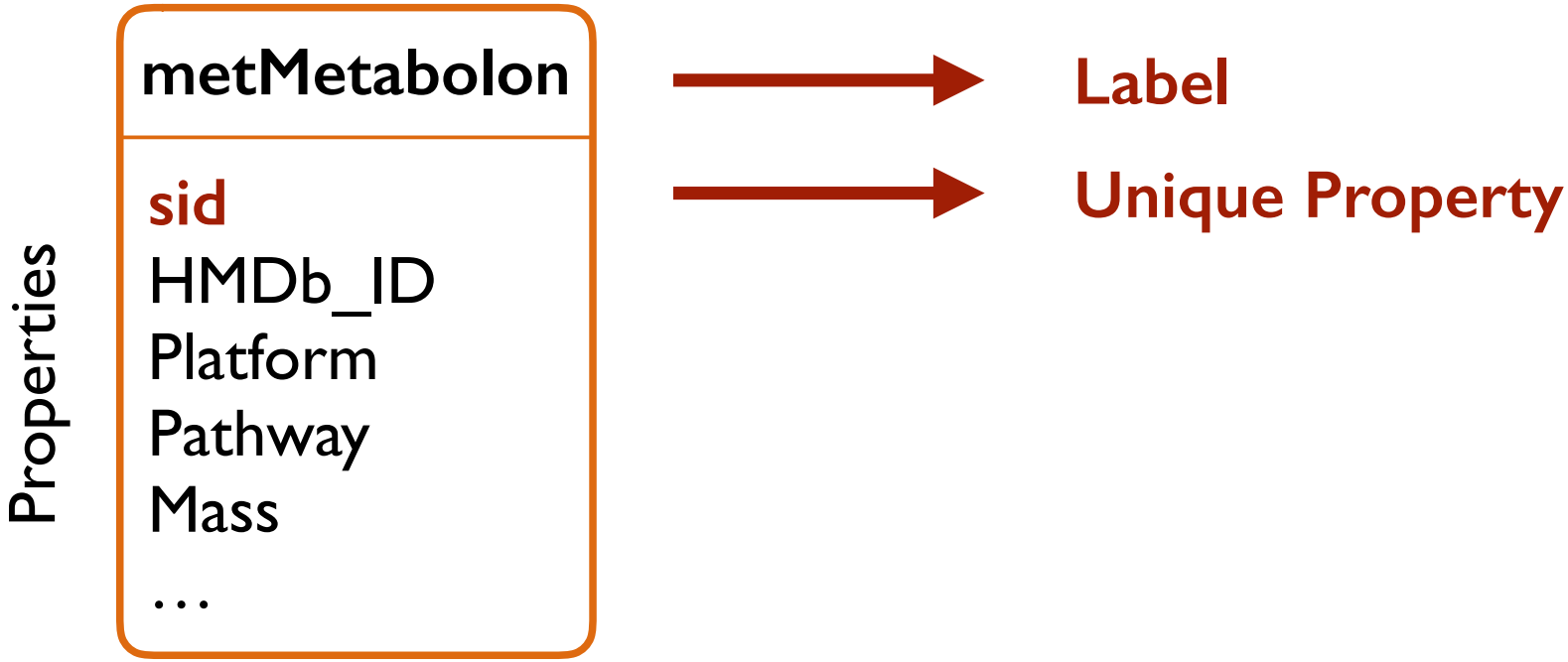
metMetabolon

sid
HMDb_ID
Platform
Pathway
Mass
…

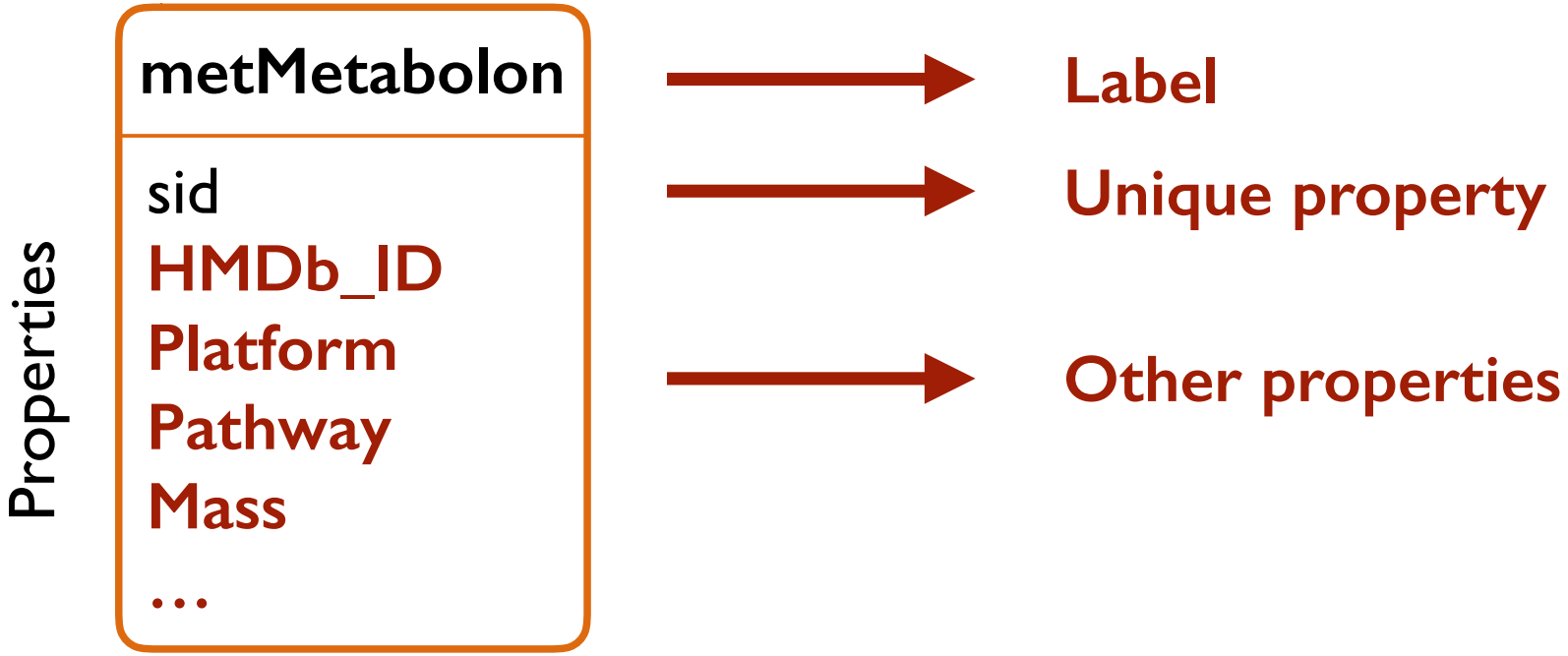
Neo4j - Nodes


metMetabolon
sid
HMDb_ID
Platform
Pathway
Mass
…
Properties
Label

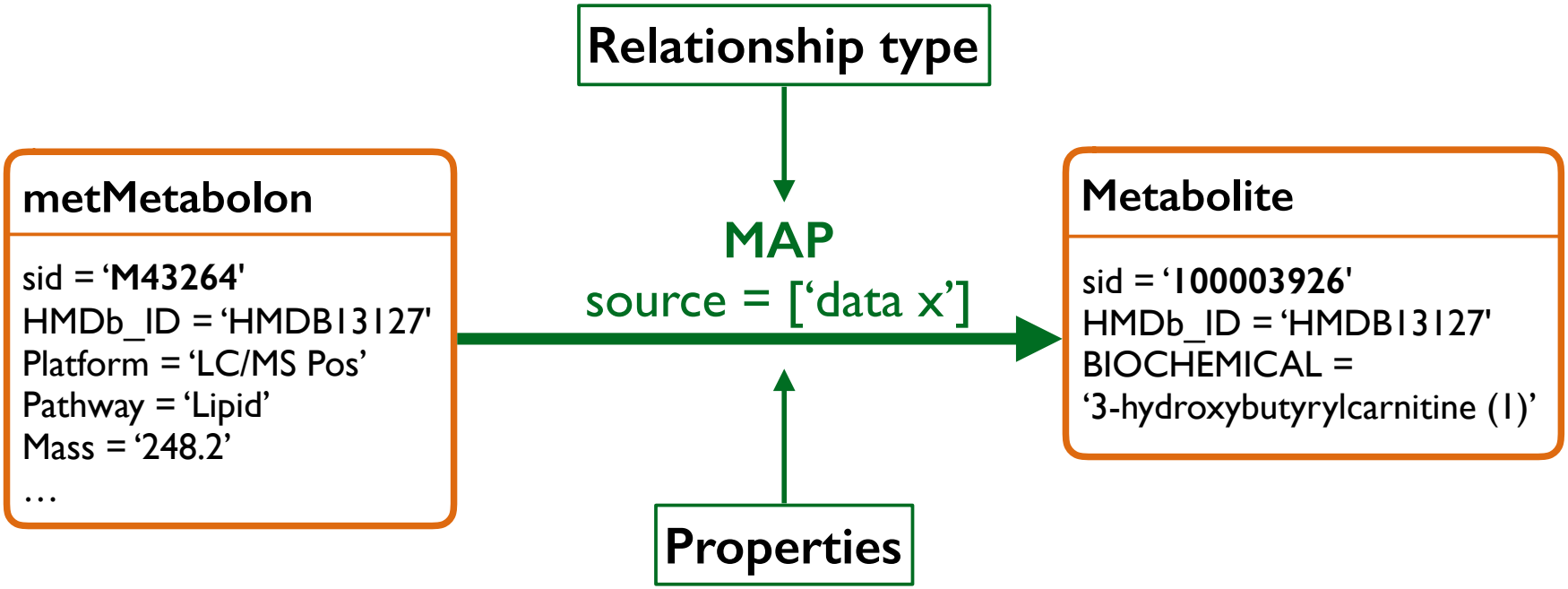
Neo4j - Nodes


metMetabolon
sid
HMDb_ID
Platform
Pathway
Mass
...
Properties
Label
Unique Property

Neo4j - Nodes


Properties
metMetabolon
sid
HMDb_ID
Platform
Pathway
Mass
...
Label
Unique property
Other properties

Neo4j - Relationships



➡ relationships always have a direction
➡ direction can be ignored in queries

Cypher

- declarative graph query language
- uses patterns to describe graph data
- can be used for querying and updating
- queries built up using various clauses

➡ familiar, SQL-like

➡ can be chained together

➡ intermediate results will be context for next clause

Queries - Simple

Match a **specific node** and return it:

variable label property

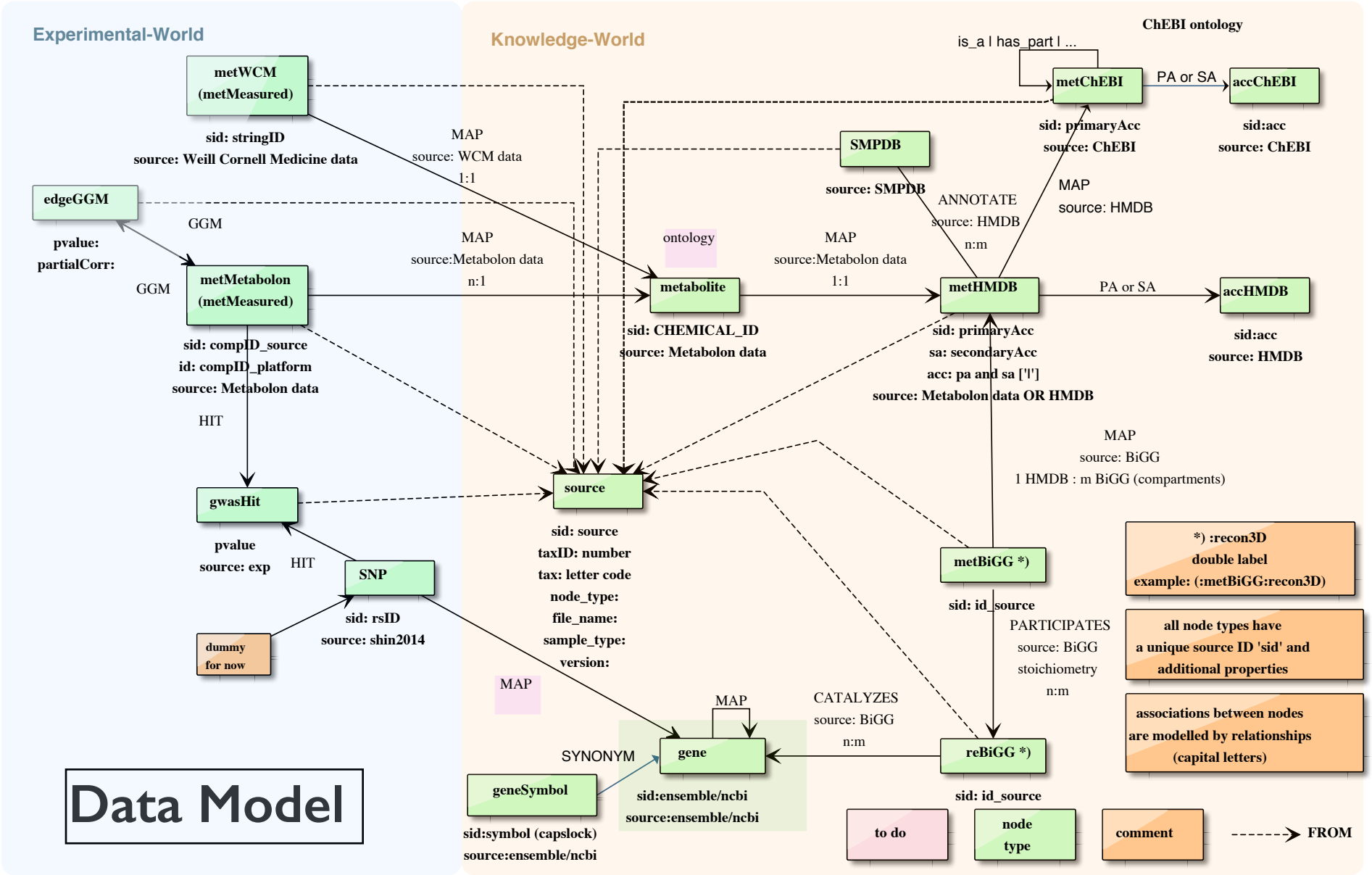
```
MATCH (m : metabolite {sid:"1021"})
RETURN m
```

Table with **all reactions** that a **metabolite participates** in:

relationship type any node

```
MATCH (m)-[:PARTICIPATES]-(r)
RETURN m.sid AS metabo, collect(r.sid)
```

only value aggregation function


Experimental-World
Knowledge-World
ChEBI ontology
is_a I has_part I ...
metWCM
(metMeasured)
sid: stringID
source: Weill Cornell Medicine data
MAP
source: WCM data
1:1
edgeGGM
pvalue:
partialCorr:
GGM
GGM
metMetabolon
(metMeasured)
sid: compID_source
id: compID_platform
source: Metabolon data
MAP
source:Metabolon data
n:1
ontology
metabolite
sid: CHEMICAL_ID
source: Metabolon data
MAP
source:Metabolon data
1:1
SMPDB
source: SMPDB
ANNOTATE
source: HMDB
n:m
MAP
source: HMDB
metChEBI
sid: primaryAcc
source: ChEBI
PA or SA
accChEBI
sid:acc
source: ChEBI
metHMDB
sid: primaryAcc
sa: secondaryAcc
acc: pa and sa ['|']
source: Metabolon data OR HMDB
PA or SA
accHMDB
sid:acc
source: HMDB
HIT
gwasHit
pvalue
source: exp
HIT
SNP
sid: rsID
source: shin2014
dummy
for now
source
sid: source
taxID: number
tax: letter code
node_type:
file_name:
sample_type:
version:
MAP
source: BiGG
1 HMDB : m BiGG (compartments)
metBiGG *)
sid: id_source
PARTICIPATES
source: BiGG
stoichiometry
n:m
*) :recon3D
double label
example: (:metBiGG:recon3D)
all node types have
a unique source ID 'sid' and
additional properties
associations between nodes
are modelled by relationships
(capital letters)
MAP
MAP
CATALYZES
source: BiGG
n:m
SYNONYM
gene
sid:ensemble/ncbi
source:ensemble/ncbi
geneSymbol
sid:symbol (capslock)
source:ensemble/ncbi
reBiGG *)
sid: id_source
to do
node
type
comment
FROM
Data Model

Git Repository

```
.
|-- code
|   |-- scripts
|   `-- setup
|-- data
|   |-- BiGGModel
|   |-- GGM
|   |-- SMPDB
|   |-- SNP
|   |-- genes
|   |-- gwasHit
|   |-- metChEBI
|   |-- metHMDB
|   |-- metMetabolon
|   |-- metWCM
|   `-- recon3d
`-- misc
```

https://gitlab.com/sysdiab/neo4j

1. code
2. data
3. misc

Git Repository

```
.
|-- code
|    |-- scripts
|    `-- setup
|-- data
|    |-- BiGGModel
|    |-- GGM
|    |-- SMPDB
|    |-- SNP
|    |-- genes
|    |-- gwasHit
|    |-- metChEBI
|    |-- metHMDB
|    |-- metMetabolon
|    |-- metWCM
|    `-- recon3d
`-- misc
```

1. code

```
code/setup
|-- README.md
|-- setup.R
`-- toLoad
    |-- parseMetabolonFile.R
    `-- setupFunctions.R
```

Git Repository

```
.
|-- code
|    |-- scripts
|    `-- setup
|-- data
|    |-- BiGGModel
|    |-- GGM
|    |-- SMPDB
|    |-- SNP
|    |-- genes
|    |-- gwasHit
|    |-- metChEBI
|    |-- metHMDB
|    |-- metMetabolon
|    |-- metWCM
|    `-- recon3d
`-- misc
```

2. data

- **File types:**
  - .xlsx
  - .csv
- **Info sheet:**
  - Properties for source node
  - CSV: file with _info.csv
  - XLSX: additional sheet "info"

Git Repository

```
.
|-- code
|    |-- scripts
|    `-- setup
|-- data
|    |-- BiGGModel
|    |-- GGM
|    |-- SMPDB
|    |-- SNP
|    |-- genes
|    |-- gwasHit
|    |-- metChEBI
|    |-- metHMDB
|    |-- metMetabolon
|    |-- metWCM
|    `-- recon3d
`-- misc
```

2. data

taxID	9606
tax	hsa
sid	HELM-16-16ML+ CDT (161123) liver
file_name	HELM-16-16ML+ CDT (161123)
node_type	metMetabolon
sample_type	liver
study	HELM-16-16ML+

- **Info sheet:**
  - Properties for source node
  - CSV: file with _info.csv
  - XLSX: additional sheet "info"

Git Repository

```
.
|-- code
|    |-- scripts
|    `-- setup
|-- data
|    |-- BiGGModel
|    |-- GGM
|    |-- SMPDB
|    |-- SNP
|    |-- genes
|    |-- gwasHit
|    |-- metChEBI
|    |-- metHMDB
|    |-- metMetabolon
|    |-- metWCM
|    `-- recon3d
`-- misc
```

2. data

```
data/gwasHit
`-- 9606_hsa
    |-- snp_metabolite_assocs_shin2014.csv
    `-- snp_metabolite_assocs_shin2014_info.csv
```

Git Repository

```
.
|-- code
|    |-- scripts
|    `-- setup
|-- data
|    |-- BiGGModel
|    |-- GGM
|    |-- SMPDB
|    |-- SNP
|    |-- genes
|    |-- gwasHit
|    |-- metChEBI
|    |-- metHMDB
|    |-- metMetabolon
|    |-- metWCM
|    `-- recon3d
`-- misc
```

2. data

```
data/gwasHit
`-- 9606_hsa
    |-- snp_metabolite_assocs_shin2014.csv
    `-- snp_metabolite_assocs_shin2014_info.csv
```

```
metMetabolon
|-- 10090_mmu
|    |-- HELM-14-15ML+CDT\ \ (160202).xlsx
|    |-- HELM-16-16ML+\ CDT\ (161123)\ WAT.xlsx
|    |-- HELM-16-16ML+\ CDT\ (161123)\ liver.xlsx
|    `-- NonTargeted_Test\ sample_results_HELM-13-1
`-- 9606_hsa
    |-- CORN-0301-09VWBL(110819).xlsx
    |-- CORN-0302-09VWBL\ (120126).xlsx
    |-- CORN-0402-11VWBL(120824).xlsx
    |-- CORN-0802-13MLBl\ lung\ cancer\ copy_fixed
    `-- Karpas-20422-20met_alt.xlsx
```

Git Repository - Data - README

FILE FORMAT CONVENTIONS:

DATA_TYPE	FORMAT	INFO	SPECIAL_REMARKS
metMetabolon	xlsx	yes	2 sheets (first named info)
metWCM (dummy)	xlsx	yes	must have an "id" and "hmdb" column
GGM	xlsx	yes	2 sheets: "info" and "GGM list": contains in order (metabolite a, metabolite b, partialCorr, pvalue)
gwasHit	csv	yes	column names: rsID, metabolite, pvalue
SNP (dummy)	csv	-	column names: sid, chr, position

INFO SHEET:

DATA_TYPE	INFO	FORMAT	ENTRIES
metMetabolon	yes	xlsx	sid, file_name, node_type, tax, taxID
metWCM (dummy)	yes	xlsx	sid, file_name, node_type, tax, taxID
GGM	yes	xlsx	sid, tax, taxID, file_name, node_type, metMetabolon_id
gwasHit	yes	csv	sid, tax, taxID, file_name, node_type

Neo4j Servers

- **Production Server:**
  - stable, read-only
  - no password
  - http://dzdcon1.helmholtz-muenchen.de:6464/browser/

- **Development Server:**
  - password protected
  - http://dzdcon1.helmholtz-muenchen.de:9494/browser/

➡ both running Neo4j version 3.3.0

Running setup.R

- clone git repository
  - set-url to git@gitlab.com:sysdiab/neo4j.git (not https)
  - not prompted for your username/password

- if Neo4j is running on <u>different machine/server</u>
  - ssh connection without password

Code

I) Paths

II) Setup

- setwd()
- load scripts from toLoad directory
- create log files
- connect to db
- define unique constraints (“sid”)

III) SCP

- git pull local & remote

IV) Fill databank

Code

I) Paths

```
#################
## PATHS
#################
git <- "~/Documents/ICB/neo4j/" # path to git repository
log_path <- "~/Desktop/"
db_addr <- "http://dzdcon1:9494/db/data/"
## only required if Neo4j is running on different server
neo4j_server <- "ssh alex@dzdcon1"
wd_path_server <- "/home/alex/sysdiab/"
```

Code

II) Setup

- library(RNeo4j)

```
## CONNECT TO NEO4J
db = startGraph(db_addr, username = "neo4j", password = "sysdiab")
clear(db)
```

```
## define constraints of uniqueness for different node labels
nodes = c(
  "metHMDB",
  "metWCM",
  "metBiGG",
```

```
add_constr(nodes, "sid", log, db)
```

Code

III) SCP

```
##################################################################
## SCP -  only needed if neo4j on different server
##################################################################
system("git pull")
system(paste0(neo4j_server," '(cd ",wd_path_server,"; git pull)'"))
```

IV) Fill databank

- SMPDB
- HMDB
- metWCM
- metMetabolon
- …

IV) Fill databank - csv

```
for (file in getFiles("SMPDB")) {

  # read info on dataset
  source <-getInfo(file, c("sid", "file_name", "node_type", "version"), err)
  #skip file if no/incomplete info
  if (!all(is.na(source))) {
    tic(paste0("Added SMPDB file ",source["file_name"]))
    #check/add source node for dataset
    checkSource(source, db, log)
    query=paste("USING PERIODIC COMMIT 1000
                LOAD CSV WITH HEADERS FROM 'file:///",file,"' AS props
                MERGE (n:SMPDB {sid:props.sid})
                SET n += props with n
                MATCH (s:source {sid:\"",source["sid"],"\"})
                MERGE (n)-[:FROM]->(s)
                ",sep="")
    cypher(db, query) # send query
    info(log, capture.output(toc())) # log
  }
}
```

IV) Fill databank - xlsx

```
## MAP METABOLON TO METABOLITE
tx <-  newTransaction(db) # new transaction
a_ply(metabo,1,function(x) { # process df row-by-row
  query <-  paste0("
                UNWIND {properties} AS prop
                MATCH (m:metMetabolon {sid:prop.sid})
                where exists(prop.CHEMICAL_ID)
                MATCH (h:metabolite {sid:prop.CHEMICAL_ID})
                MERGE (m)-[:MAP]->(h)")
  appendCypher(tx, query, properties = x) # add query to transaction
})
logit(commit(tx),log) # send all queries
```