Challenges in Building NoSQL Databases for Natural Products Research

Maria Sorokina, Christoph Steinbeck

Friedrich-Schiller University Jena, Germany

ChemBioSys

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

About me

Cheminformatics and Computational Metabolomics
Friedrich-Schiller-University, Jena, Germany

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Steinbeck Lab:
https://cheminf.uni-jena.de

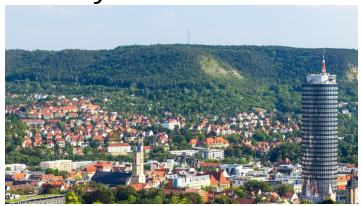
Chem- and bioinformatician

Did my studies in bioinformatics and statistics at Paris-Saclay University, France

Doctoral work at the Genoscope, Evry (Paris region), France on metabolic networks representation and finding new metabolic pathways

Now senior postdoctoral researcher at the Friedrich-Schiller University, in Jena, Germany:

- Natural Products cheminformatics (databases)
- Research Data Management for the ChemBioSys CRC
- Omics for marine diatoms

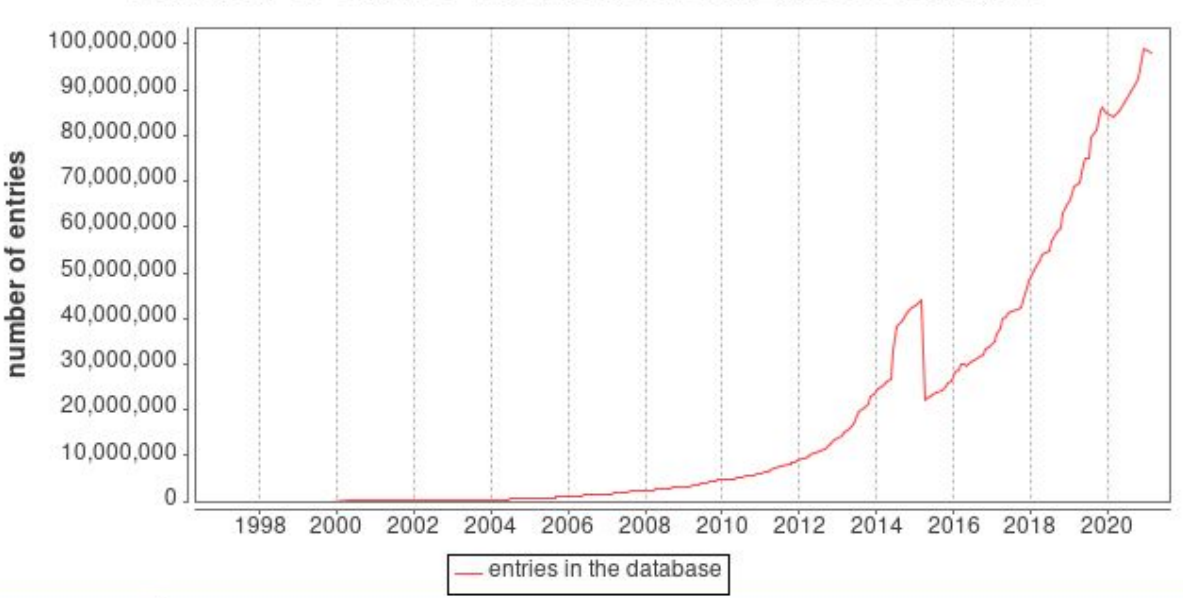


Why data storage is a challenge

Constantly increasing amount of complex data

--------> increasing need to store **increasingly complex** and **increasingly connected** data efficiently

Number of entries in UniProt/TrEMBL



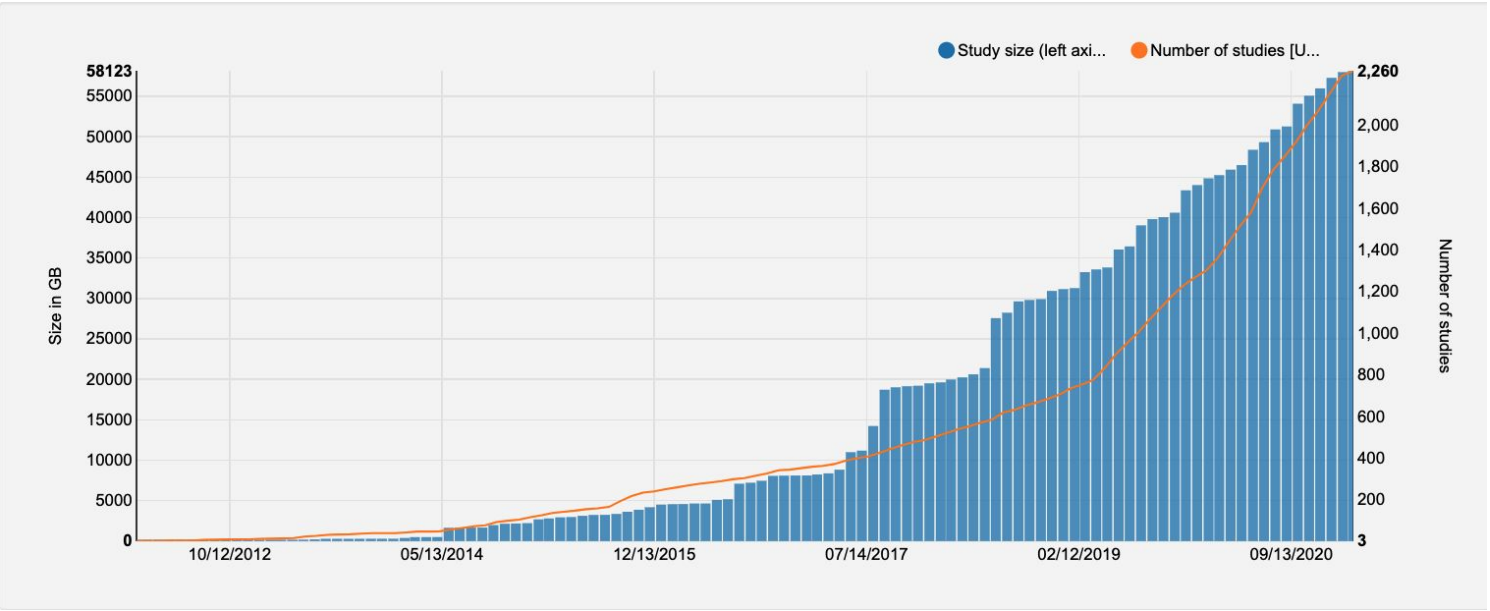
Source: https://www.uniprot.org/statistics/TrEMBL%202021_01

Why data storage is a challenge

Constantly increasing amount of complex data

--------> increasing need to store **increasingly complex** and **increasingly connected** data efficiently

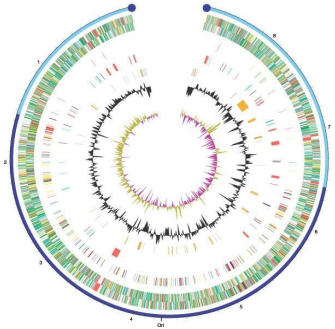
Number of entries in MetaboLights

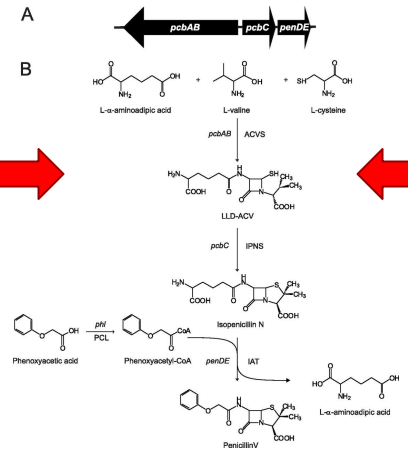


Source: https://www.ebi.ac.uk/metabolights/statistics

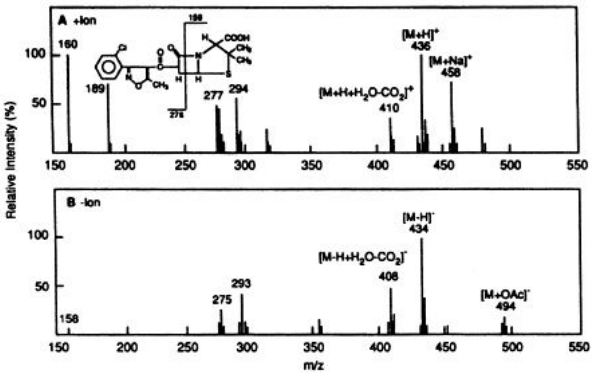
Why biological data storage is a challenge

Penicillium chrysogenum





Penicillin



Organism information & taxonomy

Assembled genome & gene functional annotations

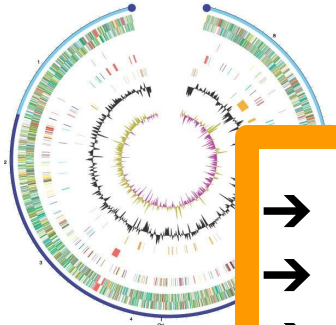
Gene clusters & biosynthetic pathways

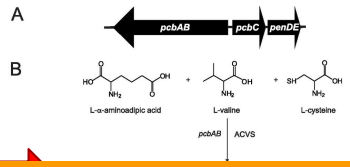
Spectral information

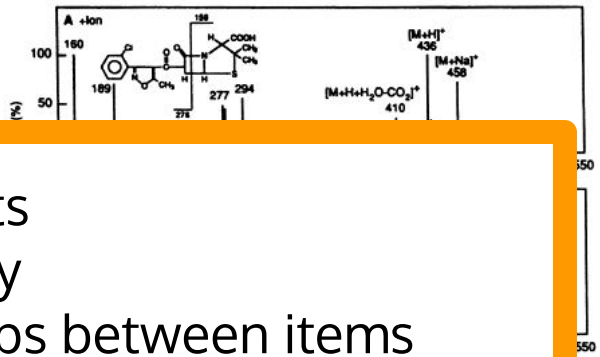
Sources: Applied and environmenatl microbiology, **Vol 87 Issue 6,**
https://aem.asm.org/content/78/19/7107/F1
https://doi.org/10.1016/S0021-9673(98)00281-7

Why biological data storage is a challenge

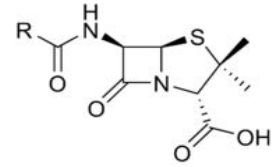
Penicillium chrysogenum







Penicillin



Organism information & taxonomy

Assembled genome & ge functional annotations

- ➔ Multiple data formats
- ➔ High data complexity
- ➔ Complex relationships between items

- ➔ **Need an efficient way of modelling and storing data**

Sources: Applied and environmenatl microbiology, **Vol 87 Issue 6,**
https://aem.asm.org/content/78/19/7107/F1
https://doi.org/10.1016/S0021-9673(98)00281-7

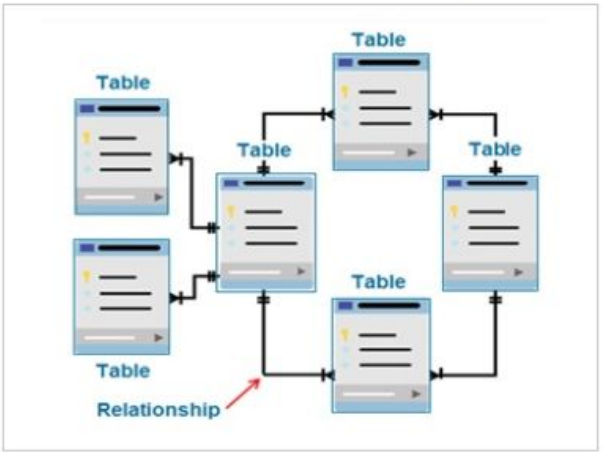
Short introduction to database management systems

SQL: Structured Query Language - main query language for relational databases

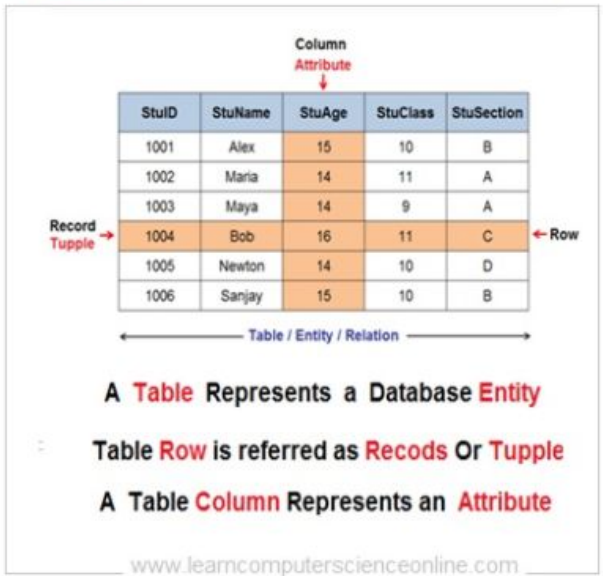
Short introduction to database management systems

SQL: Structured Query Language - main query language for relational databases

Relational Database Model



RDBMS - Relational Table Example



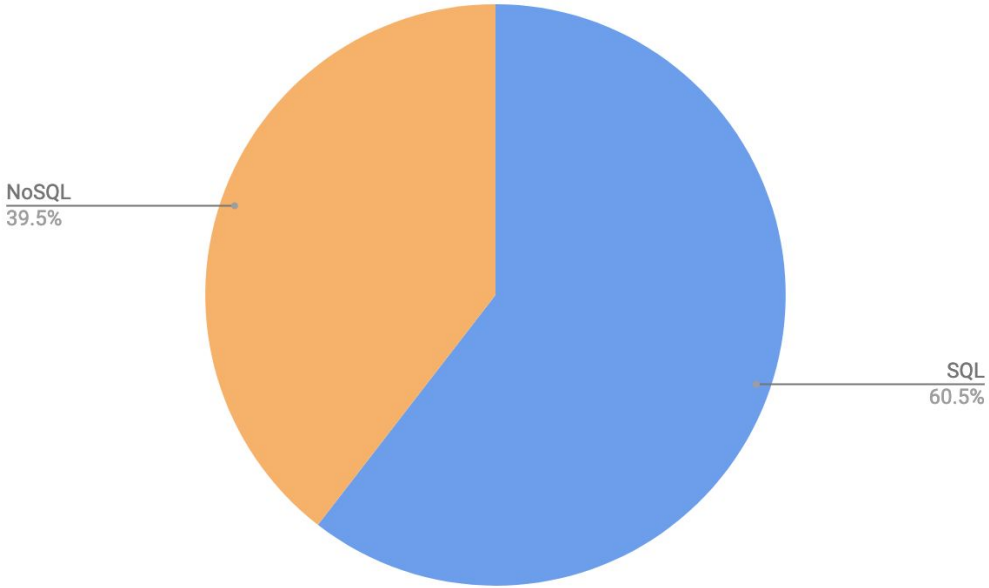
- create meaningful information by joining the tables
- joining tables allows to understand the *relationships* between the data, or how the tables connect
- good balance between flexibility and efficiency
- indexing

BUT 

- table schema is expensive to change
- not terrific with parallel writing to the same table
- limited data formats and organization

Short introduction to database management systems

In 2019:

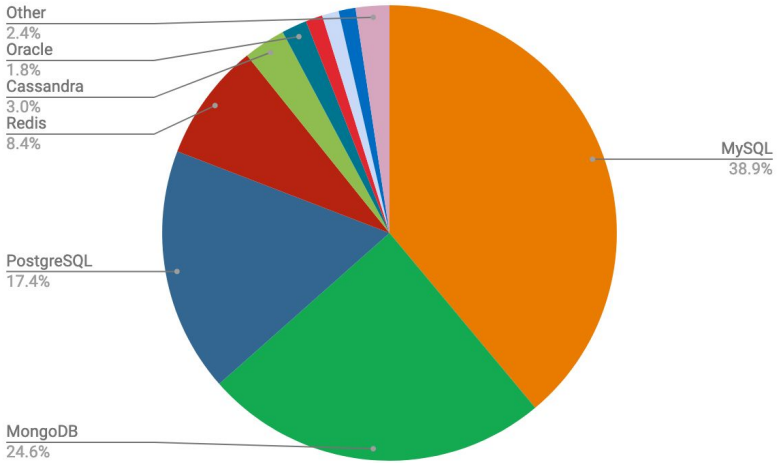


source: https://scalegrid.io/blog/2019-database-trends-sql-vs-nosql-top-databases-single-vs-multiple-database-use/

Short introduction to database management systems

SQL: Structured Query Language

noSQL: “not only SQL” rather than “not SQL”



source: https://scalegrid.io/blog/2019-database-trends-sql-vs-nosql-top-databases-single-vs-multiple-database-use/

Short introduction to database management systems

SQL: Structured Query Language

noSQL: “not only SQL” rather than “not SQL”

Less relational

More relational

Key-value DBs
Redis, Voldemort, Dynamo

Column-oriented DBs
(db-dependent, e.g. CQL)
- CassandraDB, Google’s Big Table

Document DBs
(db-dependent, e.g. mongo query language)
- MongoDB, CouchDB

Relational DBs:
(SQL)
- MySQL, PostgreSQL, MariaDB, Oracle

Graph DBs
(db-dependent, e.g. Cypher)
- Neo4j, OrientDB

noSQL database: mongoDB®

- Document database
- Mostly open (they make money with the cloud for companies)
- Mongo Query Language
- Since 2018 main storage engine: WiredTiger - extremely effective for Big Data
- Each databases is composed of “collections” (equivalent of tables)
- Each collection is composed of “documents”
- Documents are in JSON format

- Does NOT have in-built functions (yet) for chemical data (as it has for geo data)

noSQL database:

A storage engine is a software module that a database management system uses to **C**reate, **R**ead, **U**pdate and **D**elete data from a database.

- Document database
- Mostly open (they make m
- Mongo Query Language
- Since 2018 main storage engine: WiredTiger - extremely effective for Big Data
- Each databases is composed of “collections” (equivalent of tables)
- Each collection is composed of “documents”
- Documents are in JSON format

- Does NOT have in-built functions (yet) for chemical data (as it has for geo data)

noSQL database:

- Document database
- Mostly open (they make money with the cloud for companies)
- Mongo Query Language
- Since 2018 main storage engine: WiredTiger - extremely effective for Big Data
- Each databases is composed of “collections” (equivalent of tables)
- Each collection is composed of “documents”
- Documents are in JSON format

- Does NOT have in-built functions (yet) for chemical data (as it has for geo data)

Modelling data for NP: the COCONUT example

COCONUT: ColleCtion Of uNique natUral producTs

https://coconut.naturalproducts.net

In the database, 3 main collections:

- sourceNaturalProduct
- uniqueNaturalProduct
- fragment

Contains data from **56** different public data sources (version from october 2020) and **401, 624** unique “flat” molecules

Modelling data for NP: the COCONUT example

Why MongoDB?

- ➔ Structure flexibility: easy to add more complex fields if needed
- ➔ Good with big complex data
- ➔ Fuzzy text search auto-enabled
- ➔ Multiple in-build search functions (like selecting on-bits in fingerprints, which facilitates structure search)

Modelling data for NP: the COCONUT example

For each unique flat NP, a number of annotations is stored:

- Available 3D structures
- Molecular descriptors
- Traditional names, IUPAC names, synonyms, ...
- Fingerprints (enabling similarity and substructure search)
- Taxonomy
- Literature
- Geography
- Chemical classification
- Source databases (and x-references to them)

Modelling data for NP: the COCONUT example

For each unique flat NP, a number of annotations is stored:

- Available 3D structures
- Molecular descriptors
- Traditional names, IUPAC names, synonyms, ...
- Fingerprints (enabling similarity and substructure search)
- Taxonomy
- Literature
- Geography
- Chemical classification
- Source databases (and x-references to them)

Stored as simple list, enabling searchability, but doesn't keep the taxonomic tree

Modelling data for NP: the LOTUS example

LOTUS (naturaL prOducTs occUrrences databaSe) - extra-curated NP database

https://lotus.naturalproducts.net

Concept:

each NP is associated with at least one literature reference and one organism

➔ COCONUTs' simple list of taxonomies and list of references is not good enough
➔ more complex data model needed

Modelling data for NP: the LOTUS example

LotusUniqueNaturalProduct

- id
- structure representations
- taxonomyObject: {
  [{doi: '[{ taxonomyDatabase1:[{organism1,organism2,...]}, {..}]}]
- ...

Take home messages

- ★ Natural products data quantity and complexity is constantly increasing
- ★ Proper data organisation is essential at the earliest stages
- ★ MongoDB and other document-based DBMs are extremely efficient and allow without too much damage a late data model replanification

Acknowledgements

Cheminformatics and Computational Metabolomics
Friedrich-Schiller-University, Jena, Germany

Chris Steinbeck and the wonderful Caffeine group (cheminf.uni-jena.de)

ChemBioSys CRC

ChemBioSys

COLLABORATIVE RESEARCH CENTER 1127
CHEMICAL MEDIATORS IN COMPLEX BIOSYSTEMS

And the organisers of this workshop!

1. COCONUT

Stereochemistry in COCONUT:

423706 molecules with no the stereochemistry or where it was removed

- 50% only one stereocenter (but can be more in nature!)
- 23% more than one stereocenter (from the same or different DBs)
- 15.6% - truly no stereocenters
- 11.4% - have at least one stereocenter but info missing in the source database

Missing stereochemistry is a problem, as it has it's importance for molecular function.
But this can be solved mainly experimentally...