Vytváření chemických a biologických sémantických databází

Deset let vývoje a provozování Integrované databáze malých molekul

Jakub Galgonek

elixir CZECH REPUBLIC

20.11.2023

Co je cílem projektu?

Vytvořit databázi integrující různé existující databáze malých molekul a zpřístupnit je pomocí technologií Sémantického webu.

elixir CZECH REPUBLIC

Integrovaná databáze malých molekul

Language	files	blank	comment	code
Java	477	18776	1533	62244
C	88	4153	543	12803
SQL	143	1728	419	6118
Velocity Template Language	84	316	32	3727
JavaScript	23	461	66	2367
XML	16	279	24	1745
CSS/Sass	13	205	22	684
Bourne Shell	14	224	75	503
HTML	6	85	20	410
Markdown	2	66	0	327
make	9	18	0	257
SUM:	875	26311	2734	91185

elixir CZECH REPUBLIC

Co je Sémantický web?

- představen Timem Berners-Leem v roce 2001
- měl přinést posun od „web of hypertext“ k „web of data“
- “The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.”
- “The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.”

elixir CZECH REPUBLIC

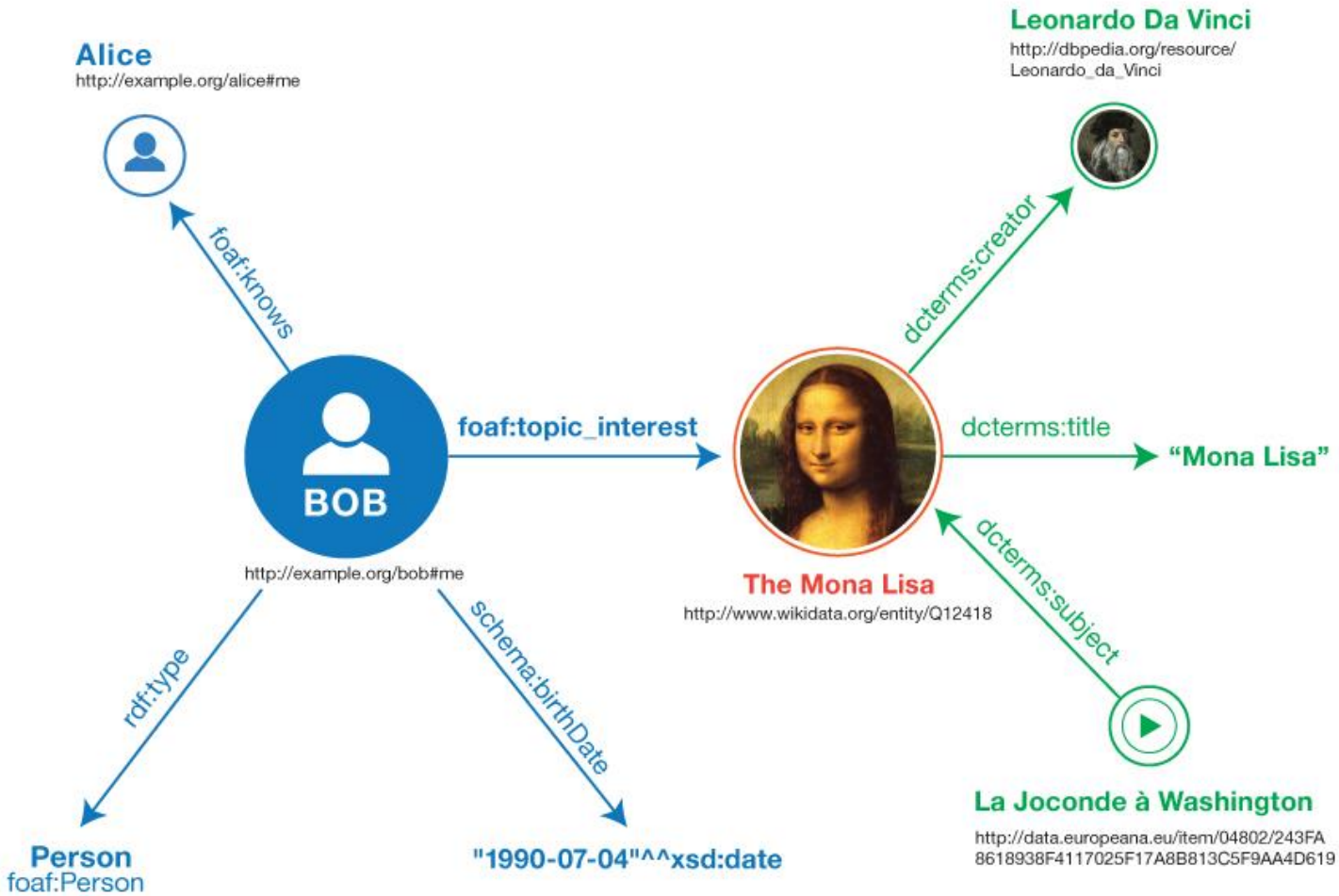
Berners-Lee, T. . et al. **The Semantic Web**. Scientific American. 2001

Resource Description Framework

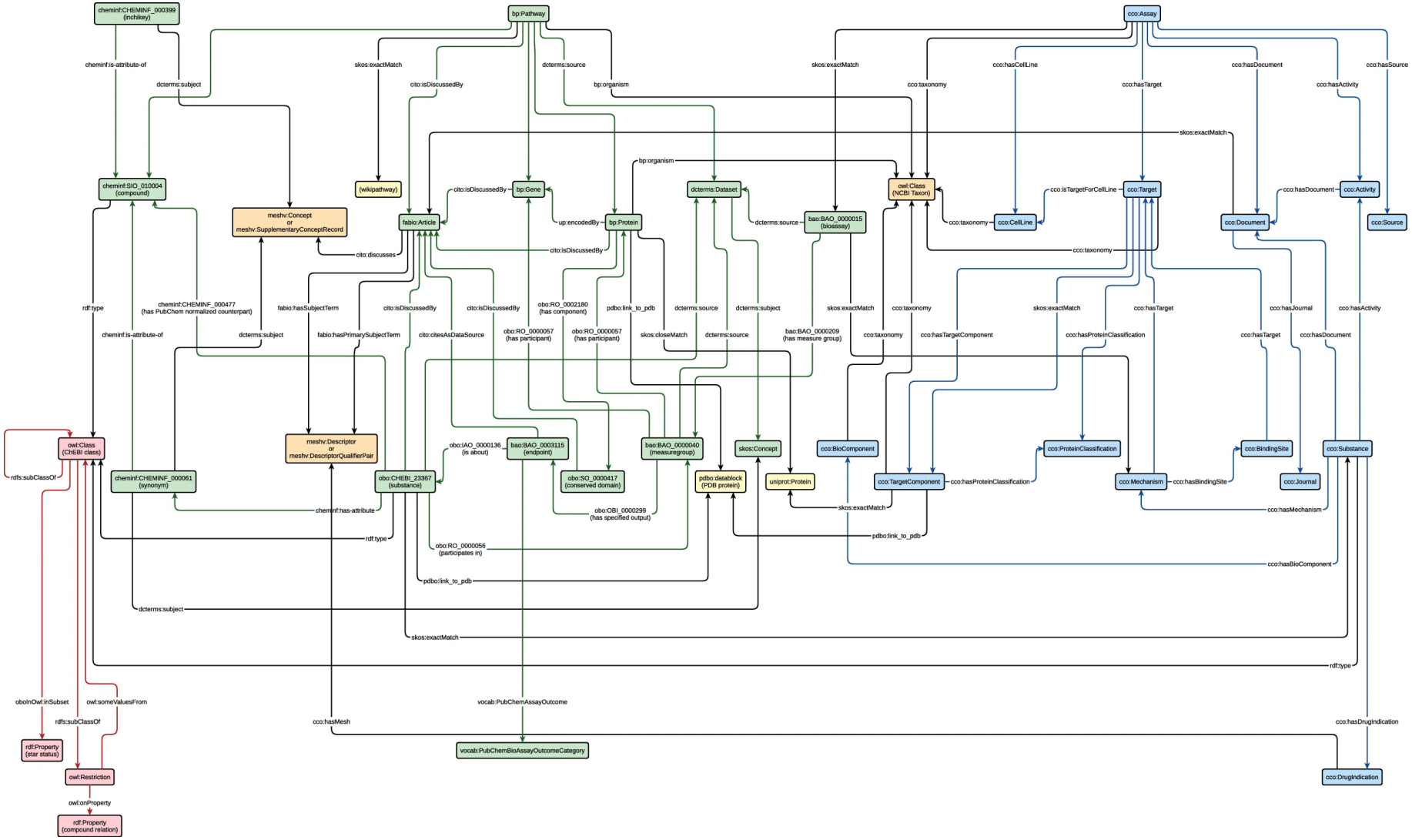
- umožňují vytvářet tvrzení o zdrojích (resources)
- pro identifikaci zdrojů a jejich vlastností se používají IRI
  - je jednoznačná
- tvrzení mají následující formu:
  - <subject> <predicate> <object> .
    - v nové verzi volitelně ještě <graph>
  - logické schéma je mnohem bližší konceptuálnímu
- pro dotazování se využívá jazyk SPARQL

elixir CZECH REPUBLIC

Resource Description Framework



https://www.w3.org/TR/rdf11-primer/example-graph-iris.jpg



elixir CZECH REPUBLIC

Cesta k projektu

- 2002-2008: magisterské studium na MFF UK
- 2004-2009: přednášky z biologie na přírodovědě
- 2008: diplomová práce o Sémantickém webu
  - téma: dotazovací jazyky pro Sémantický web
- 2008-2012: doktorát z bioinformatiky
  - téma: podobnostní vyhledávání v databázích proteinových struktur
- 2012-současnost: nástup na ÚOCHB
  - práce na IDSM (Integrated database of small molecules)

elixir CZECH REPUBLIC

Fáze I: Testujeme data (2013)

- testujeme možnost propojit různé databáze
  - ChEBI, DrugBank, PDBeChem, HMDB, NPC
- data se načítají do databáze Oracle jako RDF Semantic Graph
  - použita v předešlém projektu
- načítáme jen
  - InChI identifikátory
  - reference do databází

elixir CZECH REPUBLIC

Fáze I: Testujeme data (2013)

typ odkazů	X⇔Y					X↔Y				
	CHSI	CHsI	ChsI	Ch-I	C---	CHSI	CHsI	ChsI	Ch-I	C---
HMDB > PDBeChem	1.26	0.93	0.93	0.84	0.59	1.26	0.93	0.93	0.84	0.59
HMDB > DrugBank	7.12	5.11	4.46	1.88	1.55	7.05	5.05	4.40	1.81	1.49
HMDB > ChEBI	13.84	7.06	6.79	4.51	3.33	13.71	6.92	6.63	4.29	3.09
ChEBI > PDBeChem	9.64	8.21	6.51	4.62	1.69	9.64	8.21	6.46	4.56	1.54
ChEBI > DrugBank	**33.18**	28.85	27.95	18.96	17.85	22.99	17.77	13.74	4.54	3.25
ChEBI > HMDB	**25.80**	10.40	6.38	3.05	2.91	25.10	9.71	5.55	2.22	2.08
DrugBank > PDBeChem	**27.04**	25.49	25.17	4.04	2.11	27.02	25.47	25.13	4.00	2.01
DrugBank > ChEBI	18.33	15.08	13.98	3.13	1.91	18.10	14.85	13.69	2.84	1.57
NPC > DrugBank	23.26	13.13	12.90	6.90	5.63	22.73	12.45	11.25	4.65	3.15

elixir CZECH REPUBLIC

Fáze I: Testujeme data (2013)

- **závěr:** nelze se plně spolehnout ani na manuální ani na automaticky generované reference; problém s verzováním
  - rozhodli jsme se data neslučovat, ale pouze propojovat
- **zajímavost:** chlór měnící se na uhlík v PDBeChem

elixir CZECH REPUBLIC

Galgonek, J., Vondrášek, J. **On InChI and evaluating the quality of cross-reference links**. *J Cheminform* **6**, 15 (2014). https://doi.org/10.1186/1758-2946-6-15

Fáze II: Webová aplikace (2013-2014)

- je třeba prezentovat výsledky obecného SPARQL dotazu
  - IRI identifikátor entity není příliš uživatelsky přívětivý
  - řešení by mělo být maximálně flexibilní
    - nemělo být napevno zadrátované v kódu
- **řešení:** entity jsou spojeny s šablonami definujícími, jak je generována jejich vizuální reprezentace
- vybrán šablonovací systém Apache Velocity
  - jednoduchý, snadno rozšiřitelný
  - nevýhoda: výsledek je statický

elixir CZECH REPUBLIC

Fáze II: Webová aplikace (2013-2014)

```
1.  #sparql($roles)
2.    SELECT ?ROLEREF ?NAME WHERE
3.    {
4.      $entity chebix:hasRole ?ROLEREF.
5.      ?ROLEREF chebix:chebiName ?NAME.
6.    }
7.  #end
8.
9.  <ul>
10.   #foreach($role in $roles)
11.     <li>
12.       #url($role.ROLEREF)
13.         #escapeHTML($role.NAME)
14.       #end
15.     </li>
16.   #end
17. </ul>
```

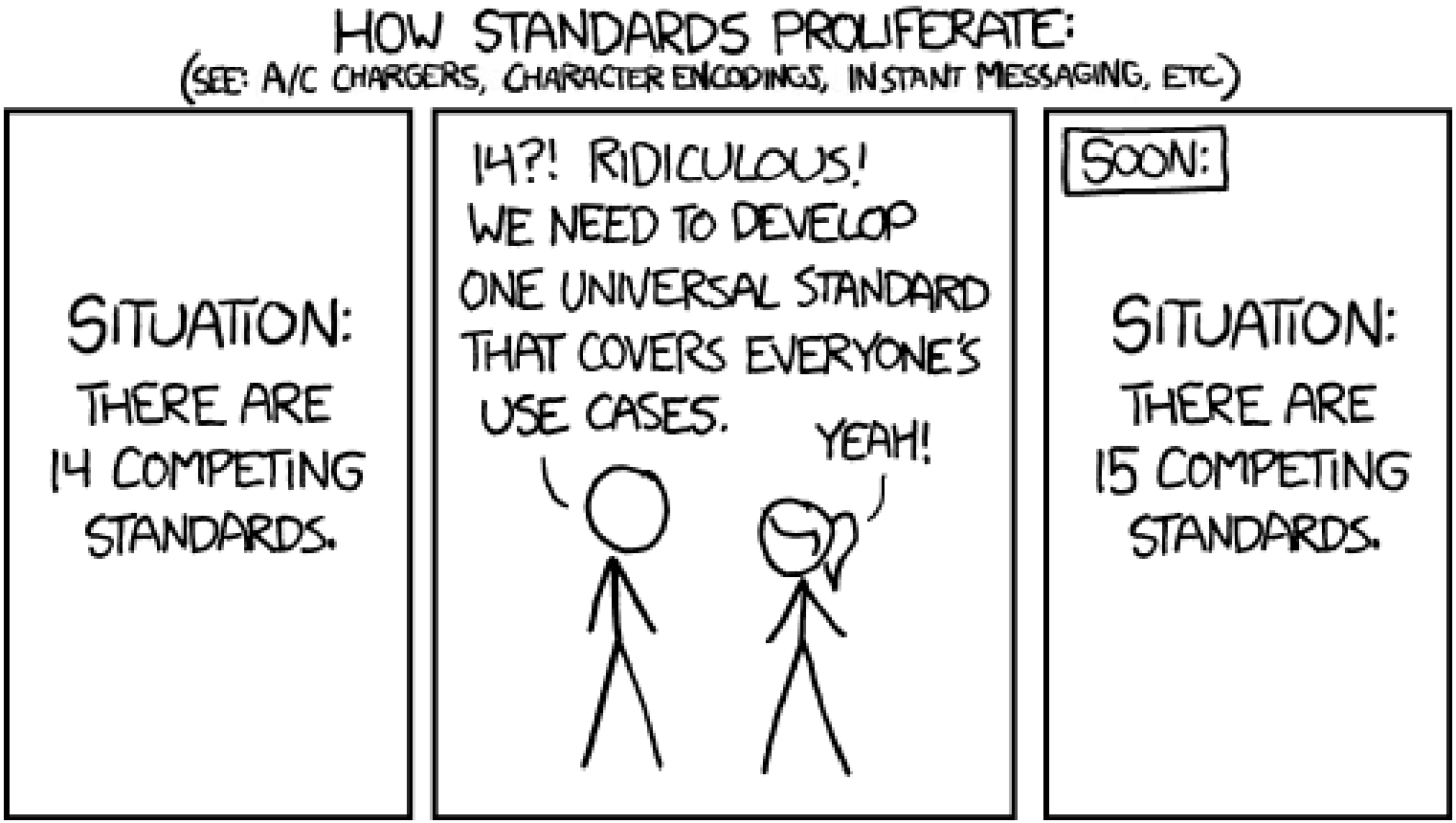
elixir CZECH REPUBLIC

Fáze II: Webová aplikace (2013-2014)

- aplikace je naprogramována pomocí Google Web Toolkitu
  - dnes již zastaralý
  - JavaScriptu se člověk stejně plně nevyhnul
  - typovou kontrolu by šlo zajistit i jinak
- samotný dotaz je psán pomocí CodeMirror
- **data:** vlastní RDF reprezentace ChEBI
  - usnadnili jsme si tím práci, ale chyba to byla už tehdy
- **závěr:** základní návrh aplikace zůstal již neměnný
- https://bioinfo.uochb.cas.cz/projects/chemRDF

elixir CZECH REPUBLIC

Fáze II: Webová aplikace (2013-2014)



https://xkcd.com/927/

Fáze III: Testování na lidech (2013)

- chemici chtějí začít něčím „jednodušším“
  - vznesen požadavek na substructure search
- použit OrChem chemistry plug-in
  - dostupný pro Oracle
  - open-source
- **závěr:** integrace OrChemu se SPARQL není přímočaře možná
  - standardní SPARQL neumožňuje volat uložené procedury

elixir CZECH REPUBLIC

Fáze IV: Problém s rychlostí (2014)

- rychlost Oracle je problémová
  - teoreticky to mohl být důsledek nedostatku našich zkušeností
  - byly zde ale i jiné problémy (cena, neúplná podpora SPARQLu)
- data přenesena do databáze Virtuoso
  - jeho rychlost je skvělá
- webová aplikace podporuje oba backendy
  - je možné přímé uživatelské srovnání
- **závěr:** je třeba se soustředit na Virtuoso
  - podpora Oracle zahozena začátkem roku 2015

elixir CZECH REPUBLIC

Fáze V: Port OrChemu pro Virtuoso (2015)

- OrChem nepodporuje Virtuoso
- Virtuoso ale podporuje Javu
- port OrChemu do Virtuosa
  - odstranění závislosti na Oracle
    - OrChem nevyužívá pokročilé možnosti Oracle databáze, pouze se potřebuje pomocí SQL dostat k datům
  - optimalizace
  - **problém:** OrChem ve skutečnosti neimplementuje stereochemii

elixir CZECH REPUBLIC

Fáze VI: OrChem z Virtuosa (2014-2015)

- neumožňuje ze SPARQLu volat funkce vracející tabulku
- umožnuje ale kombinovat SPARQL a SQL
- **řešení:** je napsat překladač ze extended SPARQLu do SQL
  - vypsán jako softwarový projekt na MFF UK
  - dokud to jde, nechává dotaz ve SPARQLu, pak přechází do SQL
  - teoreticky je schopen přeložit do SQL vše kromě basic pattern
  - **není žádoucí pozměnit syntax SPARQLu**

elixir CZECH REPUBLIC

Fáze VI: OrChem z Virtuosa (2014-2015)

```
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT * WHERE {
  [ sachem:compound ?COMPOUND;
    sachem:score ?SCORE ] sachem:similaritySearch [
          sachem:query "CC(C)Cc1ccc(cc1)C(C)C(O)=O";
          sachem:cutoff "0.8"^^xsd:double ].
}
```

elixir CZECH REPUBLIC

Fáze VI: OrChem z Virtuosa (2014-2015)

- **závěr:** jde to, ale odhaluje to mnoho chyb ve Virtuosu
  - nutnost různých workaroundů
  - komerční podpora ze strany Virtuosa je tristní
    - 8 hlášení během roku 2016
    - žádné není uzavřené

elixir CZECH REPUBLIC

Galgonek, J., Hurt, T. et al. **Advanced SPARQL querying in small molecule databases.** *J Cheminform* **8**, 31 (2016). https://doi.org/10.1186/s13321-016-0144-4

Fáze VI: OrChem z Virtuosa (2014-2015)

```
select ?X where
{
   bind ("-5"^^xsd:nonPositiveInteger AS ?X)
   filter (?X = "-3"^^xsd:nonPositiveInteger)
}
```

https://github.com/openlink/virtuoso-opensource/issues/321

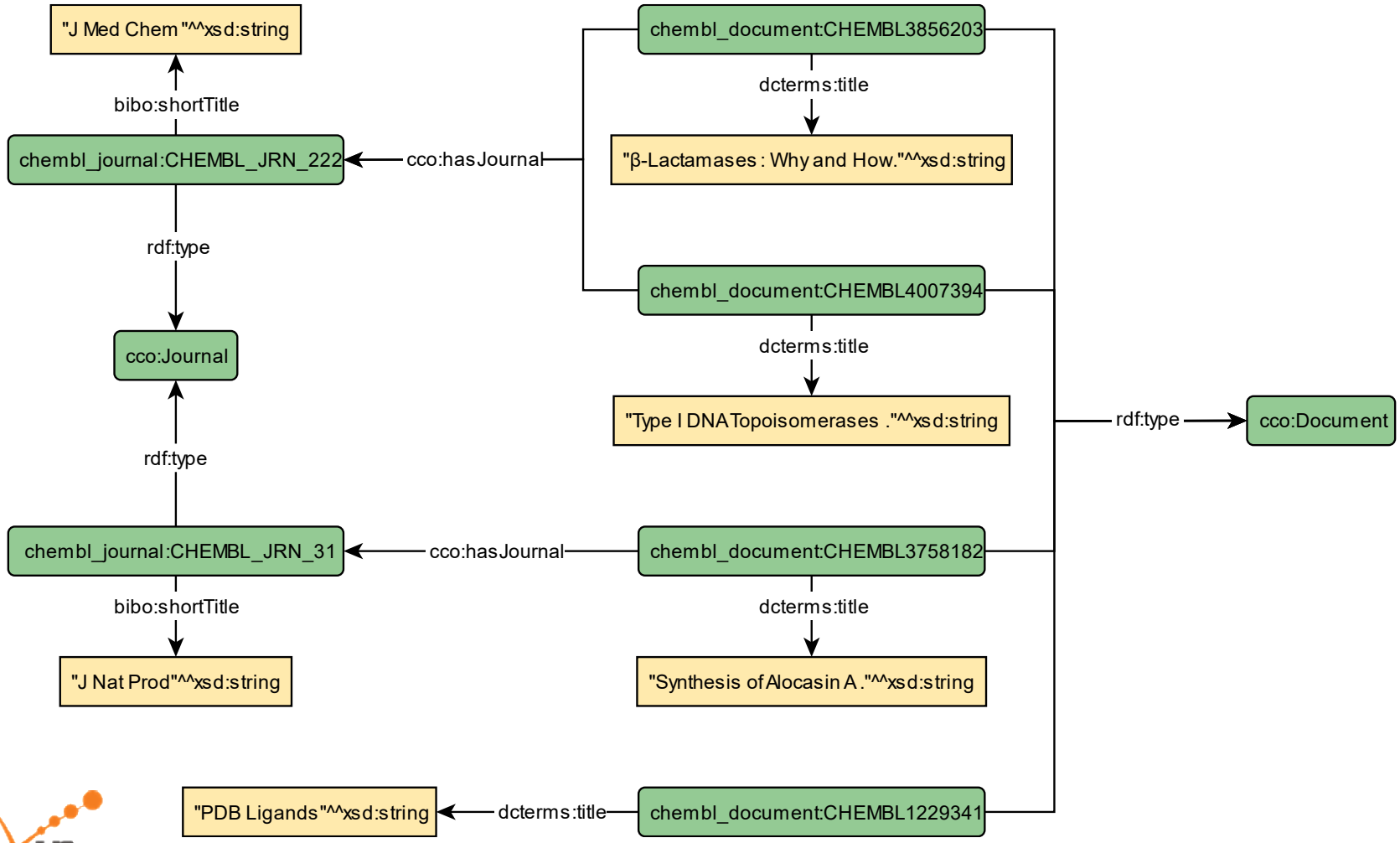
elixir CZECH REPUBLIC

Fáze VII: Big data, big problem (2016-2017)

- otestování velkých dat
  - načtena RDF databáze PubChemu
    - řádově větší než ChEBI
  - projevily se výkonnostní limity Virtuosa
- řešením je uložení dat do tabulek
  - je třeba převést sémantická data do tabulek
  - je třeba napsat mapování tabulek zpět na sémantická data

elixir CZECH REPUBLIC

Fáze VII: Big data, big problem (2016-2017)



elixir CZECH REPUBLIC

Fáze VII: Big data, big problem (2016-2017)

chembl.documents:

id	title	journal_id
1229341	PDB Ligands	NULL
3758182	Synthesis of Alocasin A.	31
3856203	β-Lactamases: Why and How.	222
4007394	Type I DNA Topoisomerases.	222

chembl.journals:

id	title
31	J Nat Prod
222	J Med Chem

elixir CZECH REPUBLIC

Fáze VII: Big data, big problem (2016-2017)

```
iri classes:
    iri:journal  = http://rdf.ebi.ac.uk/resource/chembl/journal/CHEMBL_JRN_%i
    iri:document = http://rdf.ebi.ac.uk/resource/chembl/document/CHEMBL%i

quad mappings:
    from table chembl.journals:
        iri:journal(id) rdf:type cco:Journal.
        iri:journal(id) bibo:shortTitle (short_title)^^xsd:string.

    from table chembl.documents:
        iri:document(id) rdf:type cco:Document.
        iri:document(id) dcterms:title (title)^^xsd:string.
        iri:document(id) cco:hasJournal iri:journal(journal_id).
```

elixir CZECH REPUBLIC

Fáze VII: Big data, big problem (2016-2017)

```
SELECT ?document ?title WHERE {
  ?document dcterms:title ?title.
  ?document cco:hasJournal ?journal.
  ?journal rdf:type cco:Journal.
}
```

```
SELECT
  'http://rdf.ebi.ac.uk/resource/chembl/document/CHEMBL'  tab1.document AS document,
  tab1.title AS title
FROM
  (SELECT id AS document, title AS title FROM chembl.documents) AS tab1,
  (SELECT id AS document, journal_id AS journal FROM chembl.documents
      WHERE journal_id IS NOT NULL) AS tab2,
  (SELECT id AS journal FROM chembl.journals) AS tab3
WHERE
  tab1.document = tab2.document and tab2.journal = tab3.journal;
```

elixir CZECH REPUBLIC

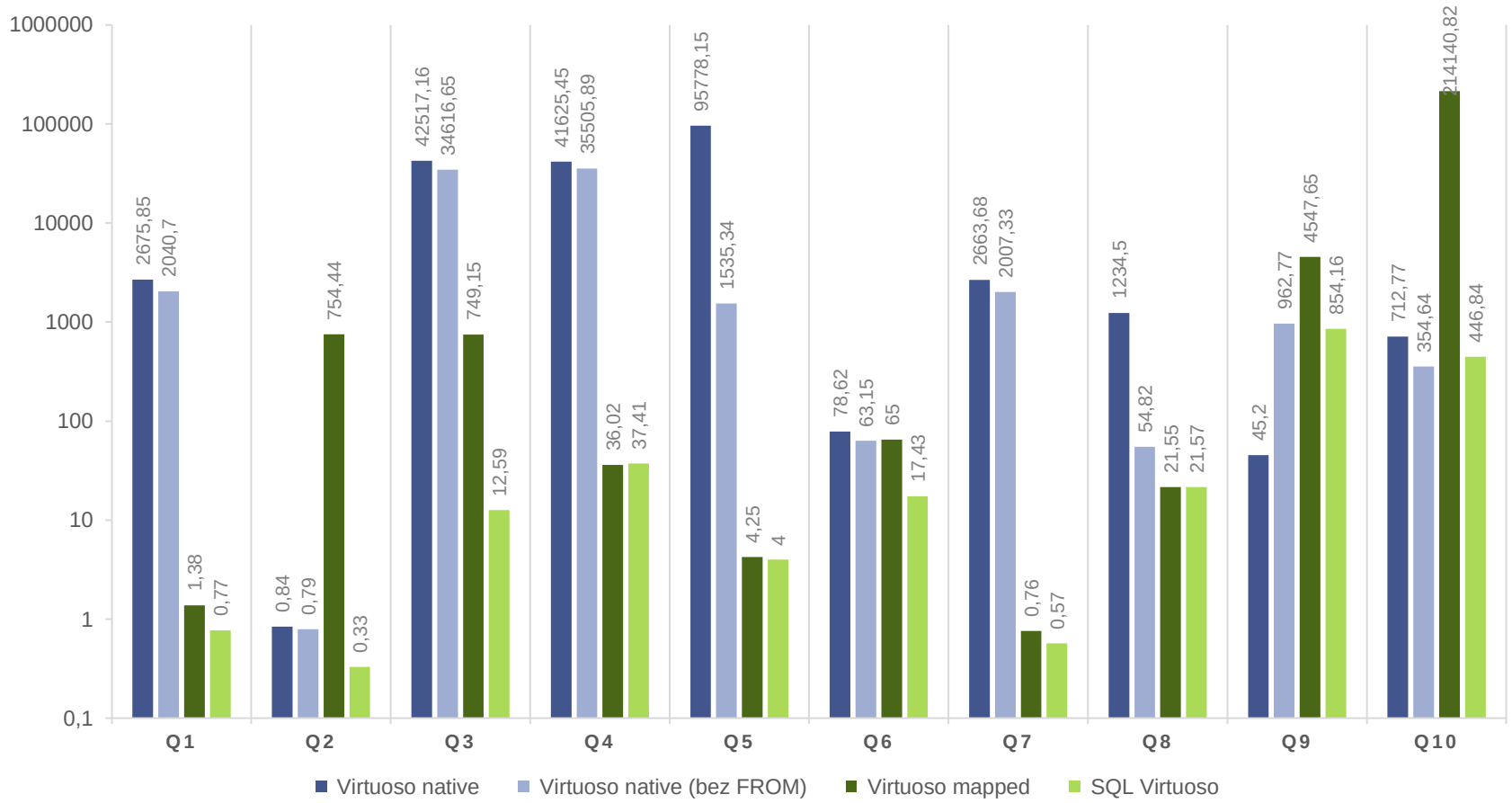
Fáze VII: Big data, big problem (2016-2017)

```
SELECT ?document ?title WHERE {
  ?document dcterms:title ?title.
  ?document cco:hasJournal ?journal.
  ?journal rdf:type cco:Journal.
}
```

```
SELECT
  'http://rdf.ebi.ac.uk/resource/chembl/document/CHEMBL'  id AS document,
  title AS title
FROM
  chembl.documents
WHERE
  journal_id IS NOT NULL;
```

elixir CZECH REPUBLIC

Fáze VII: Big data, big problem (2016-2017)



Valášek, J. **Efektivní dotazování nad databází PubChem.** (2018).
http://hdl.handle.net/10467/73981

Fáze VII: Big data, big problem (2016-2017)

- otestování velkých dat
  - načtena RDF databáze PubChemu
    - řádově větší než ChEBI
  - projevily se výkonnostní limity Virtuosa
- řešením je uložení dat do tabulek
  - je třeba převést sémantická data do tabulek
  - je třeba napsat mapování tabulek zpět na sémantická data
- **závěr:** odhalují se tím další chyby Virtuosa

elixir CZECH REPUBLIC

Fáze VIII: PostgreSQL (2017-2020)

- Virtuoso bylo definitivně zavrženo
- byl vybrán PostgreSQL
  - není sémantickou databází
  - **je třeba napsat vlastní překladač SPARQLu do SQL**
    - vycházíme z existujícího překladače
- **závěr:** vývoj překladače zabral více času, než se zdálo

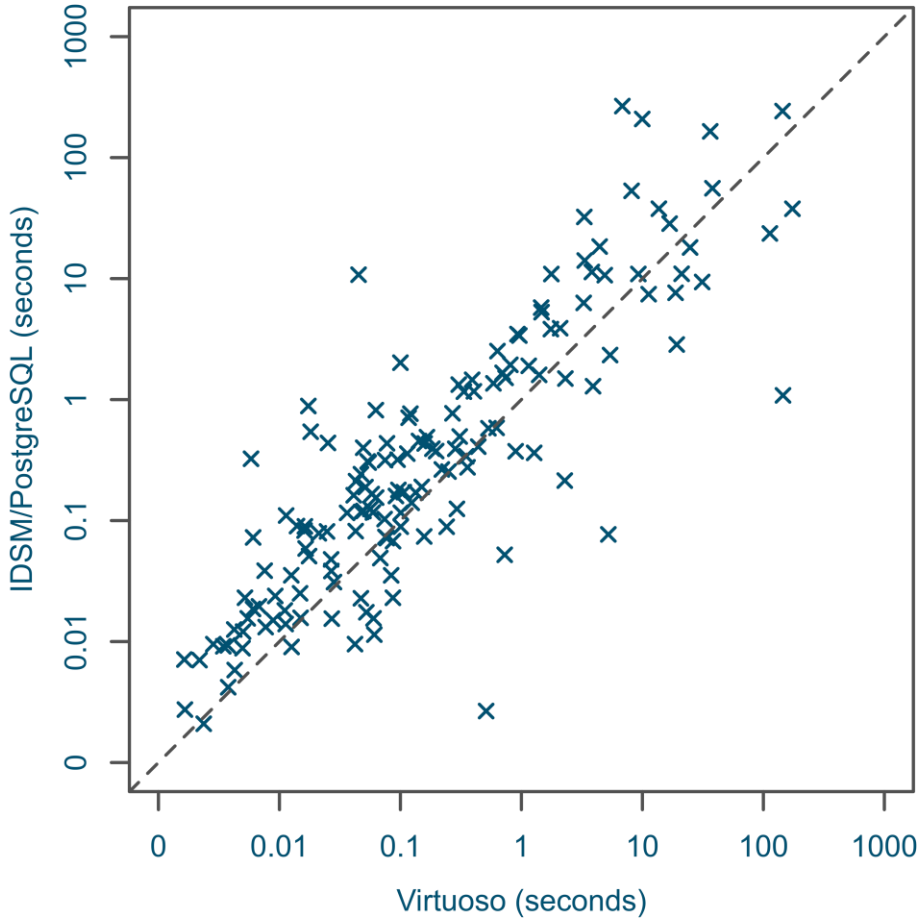
elixir CZECH REPUBLIC

Fáze VIII: PostgreSQL (2017-2020)

- je třeba naprogramovat basic pattern mattching
  - je třeba navrhnout vlastní systém mapování
- PostgreSQL nemá typ union
  - jedna proměnná ve SPARQLu může odpovídat více sloupcům v SQL
  - pro výrazy je třeba vytvořit union typ
- sémantika některých operací se liší
  - včetně aritmetických
  - některé funkce v SQL chybí
- je třeba přidat mnoho optimalizací

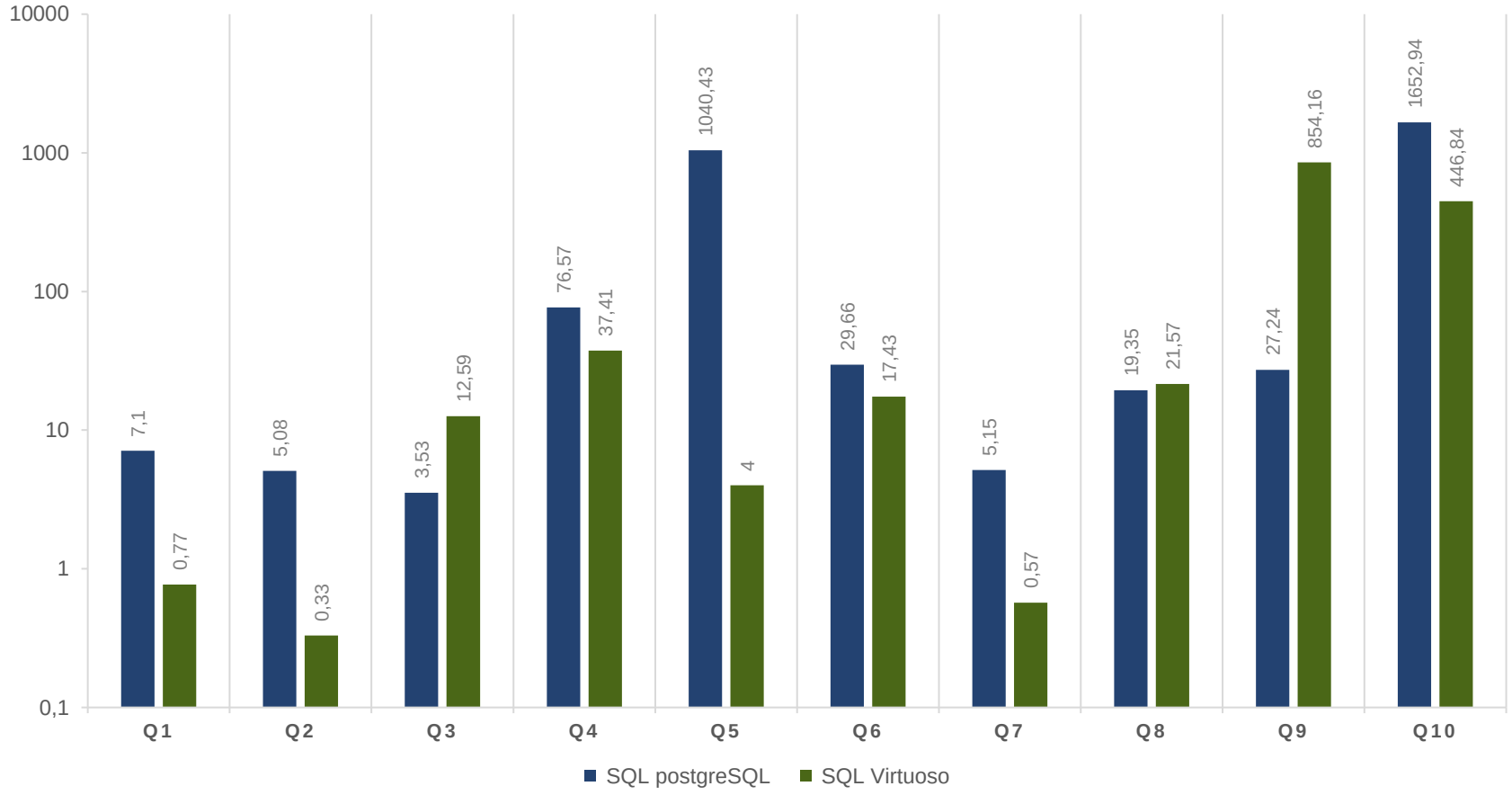
elixir CZECH REPUBLIC

Fáze VIII: PostgreSQL (2017-2020)



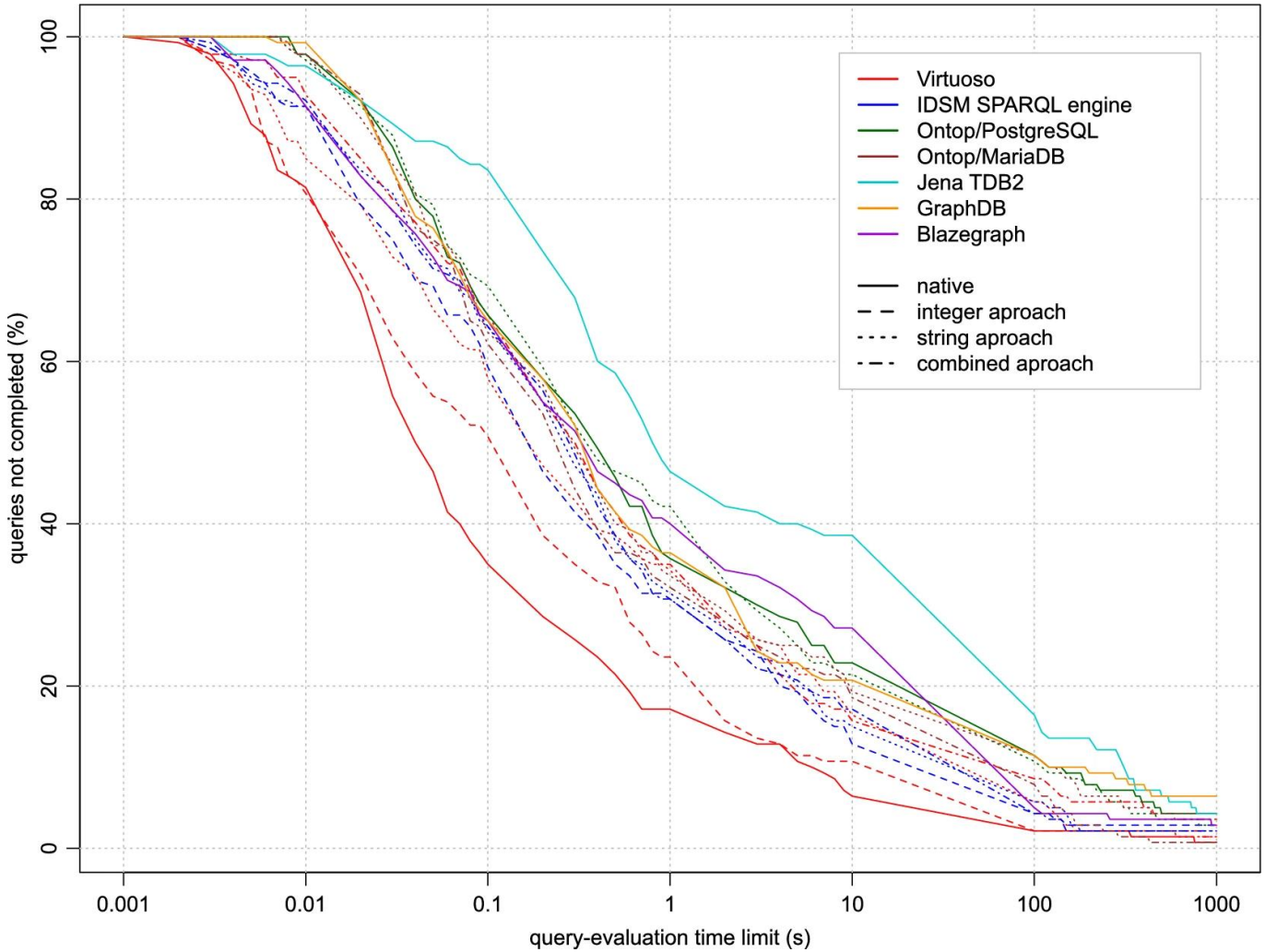
Galgonek, J. **A PostgreSQL-based RDF/SPARQL engine not only for small-molecule data.** (2019)
https://doi.org/10.7490/f1000research.1117034.1

Fáze VIII: PostgreSQL (2017-2020)



Valášek, J. **Efektivní dotazování nad databází PubChem.** (2018).
http://hdl.handle.net/10467/73981

Fáze VIII: PostgreSQL (2017-2020)



Galgonek, J. & Vondrášek, J. **A comparison of approaches to accessing existing biological and chemical relational databases via SPARQL.** *J Cheminform* **15**, 61 (2023). https://doi.org/10.1186/s13321-023-00729-5

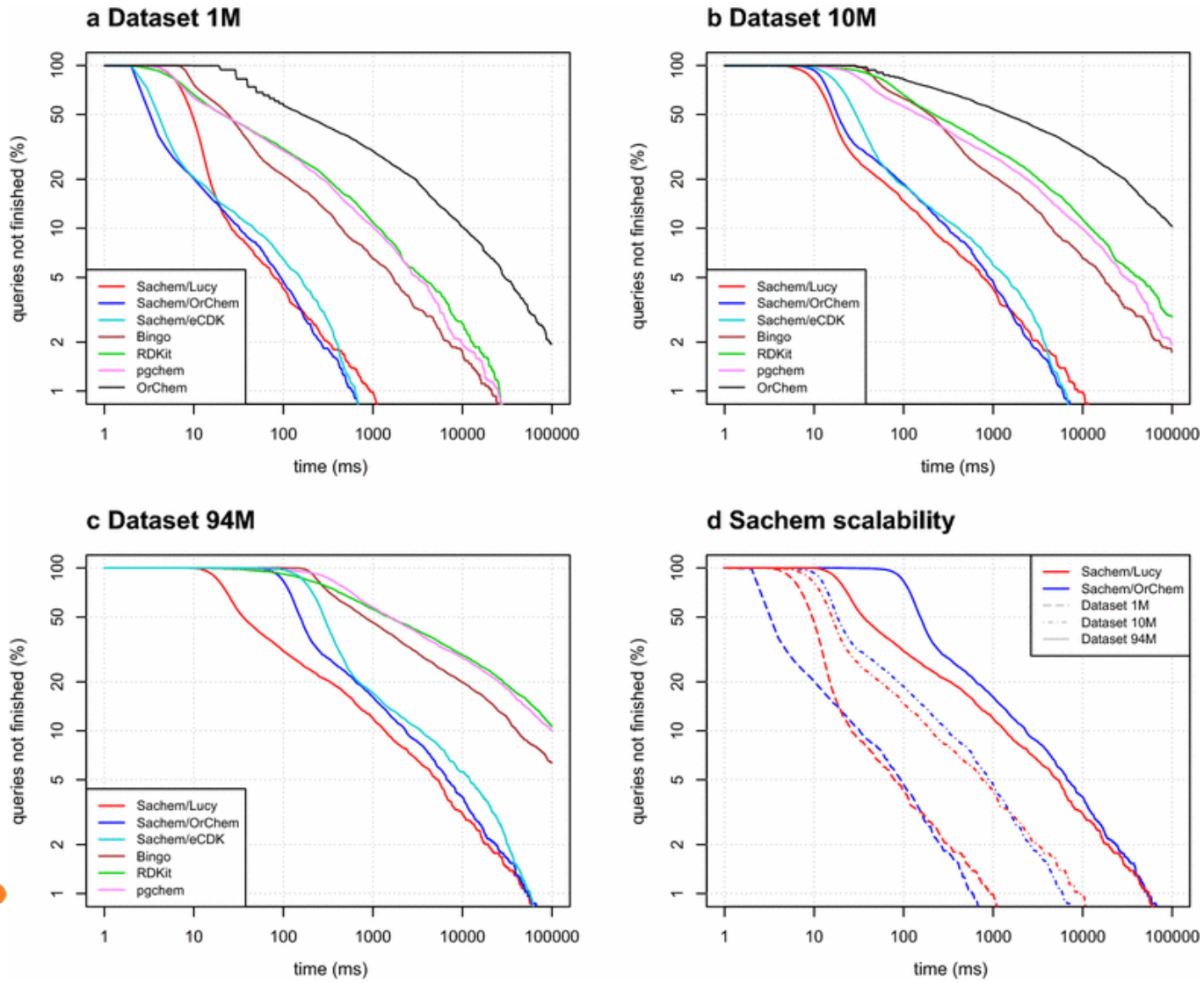
Fáze IX: Sachem přichází (2017)

- OrChem je opět portován
  - většina částí z Javy je přepisována do C
  - celkové urychlení o několik řádů
- použití fulltext indexu
  - eliminuje hodně kandidátů, ovšem stráví tím hodně času
  - přejmenování na Sachem
  - doktorandský projekt

elixir CZECH REPUBLIC

Kratochvíl, M., Vondrášek, J. & Galgonek, J. **Sachem: a chemical cartridge for high-performance substructure search.** *J Cheminform* **10**, 27 (2018). https://doi.org/10.1186/s13321-018-0282-y

Fáze IX: Sachem přichází (2017)



elixir CZECH REPUBLIC

Fáze X: IDSM jde do světa (2018-2019)

- jednoduché webové kabátky umožňující „prodat“ aktuální stav
  - překladač stále není plně hotový
  - zpřístupňujeme „jen“ substructure search skrz SPARQL
- je třeba zapracovat na automatické aktualizaci dat
- naše služba je využívána databází Rhea

elixir CZECH REPUBLIC

Kratochvíl, M., Vondrášek, J. & Galgonek, J. **Interoperable chemical structure search service.** J Cheminform **11**, 45 (2019). https://doi.org/10.1186/s13321-019-0367-2

Fáze XI: Opět Sachem … (2018-2019)

- vytváříme plugin pro Elasticsearch
  - kód z C opět přepisován do Javy
  - využití nativních metod
- plně přecházíme z Lucy na Lucene
  - držet podporu nepoužívaných metod je stojí zbytečně moc času
  - Lucy se přestává vyvíjet
  - Lucene indexy jsou menší
  - snaha využít maximum kódu z verze pro Elasticsearch

elixir CZECH REPUBLIC

Fáze XII: Konec! (2020)

- překladač je konečně ~~dokončen~~ úplný
- problém s aktualizacemi
  - struktura dat se neustále mění
- ukazuje se, že federované dotazy jsou problémové
  - přepisování dotazů
  - nedostupnost endpointů
  - nepredikovatelnost volání

Galgonek, J., Vondrášek, J. **IDSM ChemWebRDF: SPARQLing small-molecule datasets.** J Cheminform **13**, 38 (2021). https://doi.org/10.1186/s13321-021-00515-1

Fáze XII: Konec? (2020…)

- překladač je konečně ~~dokončen~~ úplný
- problém s aktualizacemi
  - struktura dat se neustále mění
- ukazuje se, že federované dotazy jsou problémové
  - přepisování dotazů
  - nedostupnost endpointů
  - nepredikovatelnost volání

Galgonek, J., Vondrášek, J. **IDSM ChemWebRDF: SPARQLing small-molecule datasets.** J Cheminform **13**, 38 (2021). https://doi.org/10.1186/s13321-021-00515-1

Fáze XII: Konec? (2020…)

```
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT * WHERE {
  [ sachem:compound ?COMPOUND;
    sachem:score ?SCORE ] sachem:similaritySearch [
          sachem:query "CC(C)Cc1ccc(cc1)C(C)C(O)=O";
          sachem:cutoff "0.8"^^xsd:double ].
}
```

elixir CZECH REPUBLIC

Fáze XII: Konec? (2020…)

```
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT * WHERE {
  _:b1 sachem:compound ?COMPOUND.
  _:b1 sachem:score ?SCORE.
  _:b1 sachem:similaritySearch _:b2.
  _:b2 sachem:query "CC(C)Cc1ccc(cc1)C(C)C(O)=O".
  _:b2 sachem:cutoff "0.8"^^xsd:double.
}
```

elixir CZECH REPUBLIC

Fáze XII: Konec? (2020…)

```
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT * WHERE {
  _:b1 sachem:compound ?COMPOUND.
  _:b1 sachem:score ?SCORE.
  _:b1 sachem:similaritySearch _:b2.
  _:b2 sachem:query "CC(C)Cc1ccc(cc1)C(C)C(O)=O".
  _:b2 sachem:cutoff 0.8.
}
```

elixir CZECH REPUBLIC

Fáze XIII: Co dál? (2023-…)

- zapracovat na lepší podpoře federovaných dotazů
- je třeba lépe vyřešit aktualizace a načítání nových dat
- portovat hlavní aplikaci z GWT do Reactu
  - oddělit tak plně backend a frontend
- je třeba přidat další optimalizace do překladače
  - hlavně rekurzivních cest
- podpora souborů s mapováním
  - aktuálně se mapování vytváří pomocí Java method

elixir CZECH REPUBLIC

Take-home message

Buďte konzervativní v tom, co posíláte,
a liberální v tom, co přijímáte.

Standard není od toho, aby se to dělalo dobře,
ale aby se to dělalo stejně.

Nepoužívejte Virtuoso.

elixir CZECH REPUBLIC