Timeline

Validation and transformation: 1 year

Database: 2 years

Evolution of databases

Key/value stores: DBM (1979)

Use of SQL (ISO 1986)

XML and XQuery

Which database?

SQLite

Since 2000

Used everywhere

existdb

Since 2006

Niche
(digital humanities)

Popularity

Questions on stackoverflow

SQLite: 93,7k	eXist-db: 534	**0.57 %**
SQL: 661k	XQuery: 5,3k	**0.80 %**

Basic architecture

SQLite

“Just” a database
(1 Mb code)

Library
(embedded)

existdb

Application platform
(163 Mb code)

Framework
(external)

Non-subjects (for now)

Data storage (update, backup)

Search interface, Web site appearance

“Traditional” query facilities

“Traditional” queries

“Find all articles published before 1967”

“Find all articles that bear the tag ‘cooking’”

Full-text search queries

“Find all Tolstoy’s books”

Everything that involves text matching

Regular expressions

TRE library: approximate search

```
cd ~/programs/dharma/repos/electronic-texts.hid/muktabodha

agrep --color -1 "(mahā)?mantreśvara" *.txt
```

Approximate parallels

See https://dharman.in

Query expansion

Goal: make queries more general, improve recall

aṅga → a(ṅ|ṃ)ga

mantras → mantr(as|aḥ|o|aś)

Inflected forms

deva → (devas|deva|devam|devena…)

⚠

Computationally expensive!

Need an automata library: OpenFST?

Difficulty: matching behaviour

Which characters are significant? Should we use a single character set?

What should we do with other characters?

What matching boundary should we choose (if any)?