On the State of the DHARMA Project

Michaël Meyer
(CESAH, CNRS)

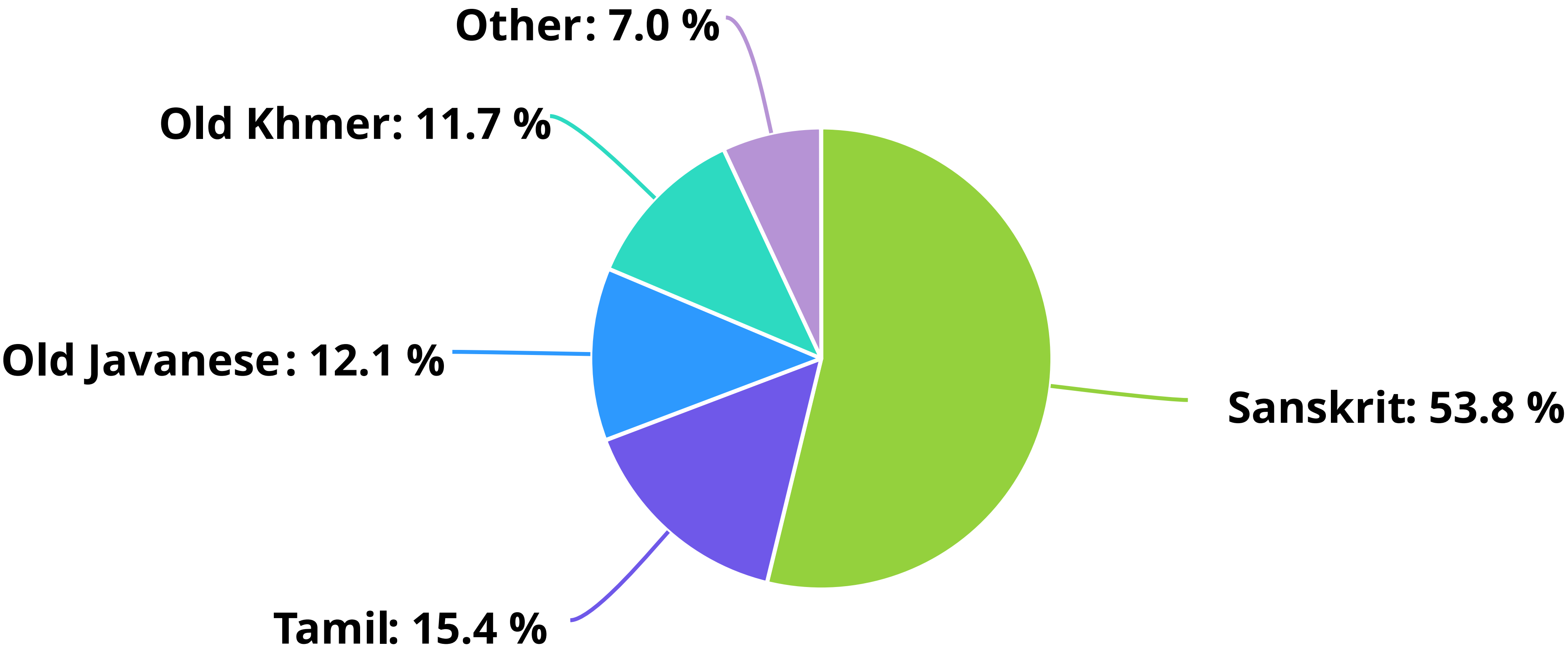
Website: dharmalekha.info

- ~50 contributors
- ~3300 texts in 13+ languages
- Inscriptions (but ~50 editions of manuscripts)

Data sources

- Original editions
- Digital editions borrowed from other projects
- Digitization of printed editions

Languages distribution



Where are we at?

- Infrastructure ✓ 🐛
- Display ✓ 🐞
- Search TODO

Infrastructure: which database?

existdb

Application platform
(163 MiB code)

Framework
(external)

SQLite

“Just” a database
(1 MiB code)

Library
(embedded)

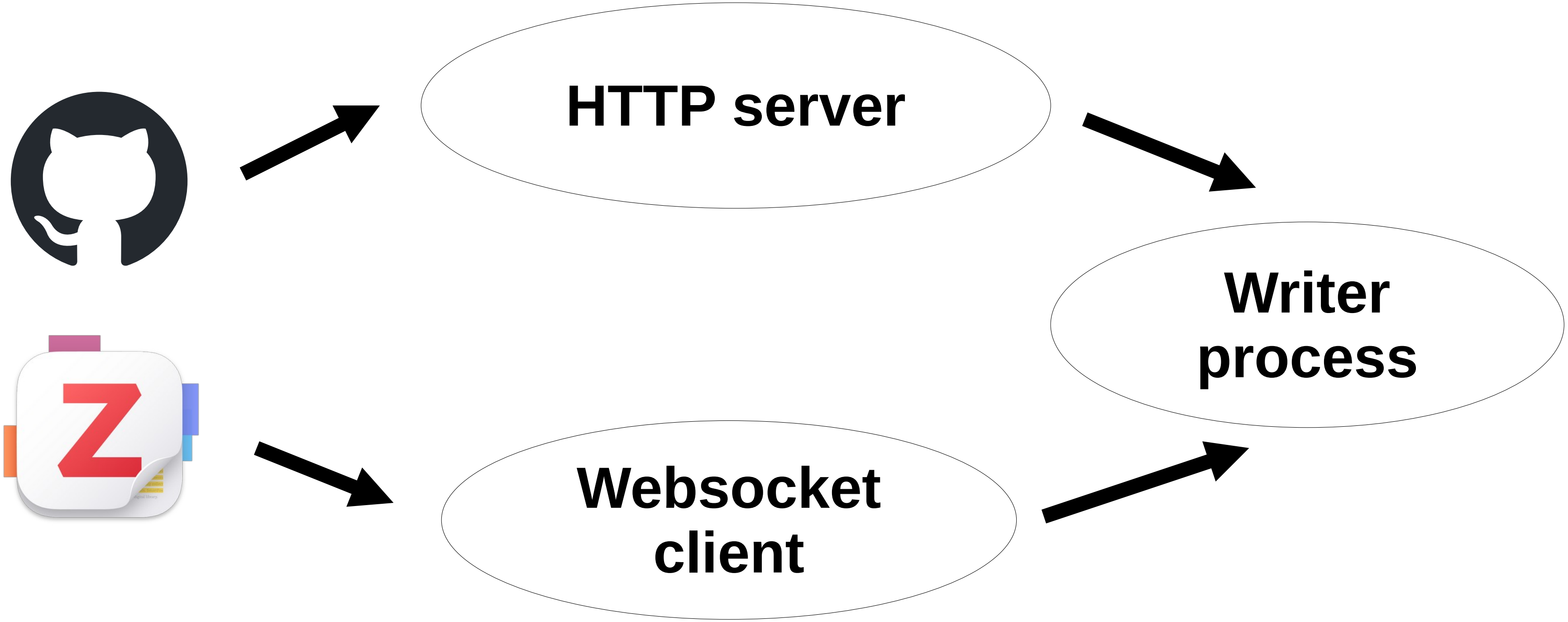
Infrastructure: concurrency

No simultaneous write transactions!

Lazy solution:

- Single writer process
- Complete isolation between database clients

Infrastructure: update system



HTTP server

Writer process

Websocket client

Display: demonstration

Text example: INSPallava00256

Search: TODO

- Write something for faceted search
- Use some external language processing tools
- Deal with source languages

Search: TRE (from Ville Laurikari)

- Approximate pattern matching (Levenshtein distance)
- Search time linear to the input size
- Constant space (fairly small)

Thank you!