Amassing and indexing a large sample of version control systems: towards the census of public source code history

This paper discusses some problems faced when trying to index mass public code and how are these problems approached by the author. The authors collected a lot of CVS hosting website at the time. Many are obsolete now such as Google Code as hosting website and Mercurial as software. After collecting the data, the authors faces some other problem regarding updating the database since collecting the data takes months of time and updating introduces many problems.

Things I would like to see discussed:

1. The paper mainly focused on data gathering and updating, but not much on indexing, I would like to have a deeper look into this.
2. The authors did not justify the completeness of the data collected.
3. The performance evaluation at many parts are very ambiguous. And the information given does not clear those questions.