

<http://promisedata.org/data>

The PROMISE repository of on-line SE data sets

Past and Present

An excessive focus on empirical results can stunt the development of innovative ideas that are, as yet, pre-experimental. However, currently, the field of software engineering is in no danger of an excess of empiricism. Hence, in 2004, Jelber Sayyad and Tim Menzies started publishing their data sets in the newly-formed PROMISE repository.

In 2008, at last update, the repository holds 88 data sets in the following areas:

- Defect Prediction (56)
- Effort Prediction (11)
- General (7)
- Model-based SE (5)
- Text Mining (9)

Why so much data? First, there is the open source effect: public code and public logs means more data sets. Second, the nature of an SE project means that once a tracking system is in place, then each new project (and each new release of each project) generates yet another data set.

Further contributions are currently being collected from the 2009 PROMISE participants, and we hope to announce 100 data sets by the time of the conference.



Future

As to the future, Susan Sontag once said "The only interesting answers are those that destroy the questions". Every repository goes through four phases.

- In phase one, everyone says "no one will give you data".
- In phase two, data arrives but no one uses it.
- In phase three, many papers are based around the data which leads to the stage four.
- In this last phase, the repository is so successful and widely used that it becomes a bias in its own right. Papers are published detailing how (e.g.) the repository is constraining further progress since the data contains systematic biases that only allow conclusions of a certain form.

Right now, PROMISE has entered phase 3 and we hope it stays there for some time. Nevertheless, it is a worthy research topic to explore limitations and restrictions on the current version of the repository and ask how might it be best changed, or torn down and started afresh.

Thanks

Finally, a note of thanks. The PROMISE steering committee wishes to thank Gregory Gay, a masters student at West Virginia University, for all his hard work in maintaining the repository.