Qualitas Corpus Clone Collection

The Qualitas Corpus Clone Collection (Collection) is part of the Qualitas Corpus (Corpus), which is a *curated* collection of software systems intended to be used for empirical studies of code artefacts. The Collection consists of data describing possible *code clones* — code fragments that are in some way similar to each other — found in most systems in the Corpus. The hope is that the accuracy of this data will be established (that is, error bounds will be provided) and, ideally, improved over time. All data provided should include its provenance — where the values came from. This will help provide some idea of how much the data can be trusted.

News

24 January 2013

The first release for the Collection is planned for 1 May 2013.

Index

Collection Catalogue	Download the collection
Structure of the Collection	Description of data
Provenance information	Citing the collection
Development status and plans	History
FAQ	Glossary

Manuscripts and Publications

- A replication and reproduction of code clone detection studies *Chen, Wang, Tempero. January 2013.*

  An unpublished paper describing the Clone Detector used to create the datasets in the first release of the Collection, and its application to part of the Corpus.

- Towards a Curated Collection of Code Clones *Tempero*. IWSC 2013.

  This is what was submitted to IWSC.

Related Information

- Yang Yuan; Yao Guo; CMCD: Count Matrix Based Code Clone Detection *APSEC 2011* The original CMCD paper

- Clone Project by the Software Engineering Group at the University of Bremen. Developers of the Rich Clone Format (RCF), an extendible data format to store and analyze code clone data.
- Stefan Bellon's Clone Reference Set

[an error occurred while processing this directive]