# Dataset for analyzing developer interactions on StackOverflow

The "dataset.xml" contains a set of posts from StackOverflow. Each post is a "row" in the dataset. Each row has the following attributes:

* Id
* PostTypeId
  + 1: Question
  + 2: Answer
* ParentID (only present if PostTypeId is 2)
* AcceptedAnswerId (only present if PostTypeId is 1)
* CreationDate
* Score
* ViewCount
* Body
* OwnerUserId (present only if the user has not been deleted; always -1 for tag wiki entries (i.e., the community user owns them))
* LastEditorUserId
* LastEditorDisplayName
* LastEditDate
* LastActivityDate (the date and time of the most recent activity on the post. For a question, this could be the post being edited, a new answer was posted, a bounty was started, etc.)
* CommunityOwnedDate (present only if post is community wikied, i.e., owned by the community user)
* ClosedDate (only present if the post is closed)
* Title (only present if PostTypeId is 1)
* Tags (only present if PostTypeId is 1)
* AnswerCount (only present if PostTypeId is 1)
* CommentCount
* FavoriteCount

Note: The StackOverflow data is a subset of the dataset used for MSR Challenge 2013 (<http://2013.msrconf.org/challenge.php>). You can cite it as follows (in bibtex format):

@INPROCEEDINGS{MSRChallenge2013,

author={Alberto Bacchelli},

title={Mining Challenge 2013: Stack Overflow},

booktitle={The 10th Working Conference on Mining Software Repositories},

year={2013},

url={\url{<http://2013.msrconf.org/challenge.php>}}

}

Latest data for StackOverflow can be found on Stack Exchange's releases of "data dumps" on archive.org (<https://archive.org/details/stackexchange>). The information is also queryable over online at the Stack Exchange Data Explorer (SEDE, <http://data.stackexchange.com/stackoverflow/queries>).