INVocD: Identifier Name Vocabulary Dataset

Simon Butler, Michel Wermelinger, Yijun Yu & Helen Sharp

Centre for Research in Computing
The Open University

May 19, 2013

The Open University

Centre for
Research in Computing

INVocD: Identifier Name Vocabulary Dataset

- Motivation:
  - Wanted a platform to build analytical tools on that makes identifier names & vocabulary directly accessible
- Dataset:
  - 60 FLOSS Java projects
  - 5,091,000 program entities, 831,000 identifier names, & 25,000 component words
  - Relational database – example SQL queries in the paper
- Uses:
  - Identifier name tokeniser development (*ECOOP '11*)
  - Part of speech tagging (*ICSM '11*)
  - Concept location (*ICSM '11*)
  - ...