Code Clones: Reconsidering Terminology

Andrew Walenstein[1]

University of Louisiana at Lafayette, Center for Advanced Computer Studies,
P.O. Box 44330, Lafayette, LA 70504-4330, U.S.A.
walenste@ieee.org

Abstract. This report discusses terminology choices and considerations relating to copied or redundant code within software systems, i.e., relating to "code clones." Inadequacies of existing terminology are raised and alternative terms are discussed.

Keywords. code clone, exact clone, near clone, clone types, accidental clone, duplicate, copy, redundant

1 Introduction

The organizers of the "Duplication, Redundancy, and Similarity in Software" (DRASIS) workshop—held at Dagstuhl, Germany in the summer of 2006—deliberately chose not to use the term "clone" in the seminar name [1]. One of the reasons is the persistent problem of what the very definition of the "clone" term is. While the term is now popular in the field, it is not at all clear that the term is universally understood in the same way. Several differences in opinion regarding what constitutes a "clone" can be found within the literature [2]. Because of this, one of the aims of the organizers was to promote renewed discussions concerning the terms and definitions used in the field [3].

During the final breakout session at DRASIS, the question was seriously raised again as to whether the term "clone" is appropriate or not, and if the scope of its use should be restricted considerably. In our opinion, by the end of the workshop there appeared to be a general consensus that the term "clone" should be semi-retired; that in the future it should be used only to refer to a more restricted concept. If indeed a conviction to rethink terminology was born then one of the main goals of the organizers was fulfilled.

If the old word use is retired or changed, however, new words or definitions need to take its place. In the final session of the workshop it was not clear to us how to fill the void created by abandoning the term "clone." This report was created to record essential aspects of the workshop discussion so that the debates and advances made there are not lost to the sands of time. What is wrong with the term "clone" and what other terms might do? What really are the issues hidden by the ill-defined "clone" term?

Although this report proposes no answers to the above questions, it records some of the discussion points as questions regarding appropriate concepts and terminology. Then, a short survey of English usage of related words is provided.

Dagstuhl Seminar Proceedings 06301
Duplication, Redundancy, and Similarity in Software
http://drops.dagstuhl.de/opus/volltexte/2007/965

We hope that this survey of terminology issues and definitions will give researchers in the field pause to think about the terms they use, and to serve as a starting point for further work.

2 What to Refer To: The Terminology Question

A suitable starting point for a discussion of possible new terms is the observation that the term "clone" frequently connotes the product of an exact duplication process. As in "The sheep Dolly is a clone of another sheep." Dolly isn't *merely similar* to some other sheep. No newspaper headlines would have trumpeted that fact. Dolly is interesting as a clone because she is genetically indistinguishable from her clone. The analogue in software is a segment of code that has been copied verbatim by some copy-paste action. This definition does not always coincide with usage in the "code clone" field. It is common within this field to talk about "clones" even if it is relatively rare to be concerned with verbatim copy-paste code segments. More likely one is dealing with a copy-paste-edit modification or even some fragment of similar code that had some independent genesis. So a better term is needed—possibly several.

It may be helpful to first have an understanding of the different contexts, issues, or debates that are raised concerning the term "clone." These include:

1. How small can clones be? Are small clones even clones? Some argue small or trivial copies of functions are not clones, perhaps because they are too small to worry about. Is a three line copy a clone? How about a one line copy? An expression?
2. Do clones have to actually have been created via copy and paste actions? Can a function be said to be a "clone" of another identical one when it was constructed completely independently?
3. Do clones need to be complete syntactic units?
4. Is a pair of similar code segments a clone pair if they are not easily refactorable? If they are not worth refactoring? If nobody wishes to refactor them [4]?
5. Is the definition of a clone universal or is it relative to some language, system, or evolution task?

Some of the above debates might properly be called "definitional" debates, not "terminological" ones. To see what we mean, consider that two particular concepts might have precise definitions, yet the terms used to denote them might not accurately denote the definitions. A single term might be used for both (aliasing), or an inappropriate or misleading term might be used to denote one of the concepts (such as calling a circle a "square" against normal convention). At this point it seems clear that there are some definitional issues to iron out [5], and that some concepts in the field might be currently more crisp than others. In this report we try to put aside definitional issues and focus solely on terminology.

The first question on terminology is understand what is being referred to. If it were just one concept (even vaguely defined), then perhaps the problem would be less severe. From the above list there seems to be are several different related concepts that are being referred to; it seems likely that this is a major factor contributing to the confusion in terminology. To try to clarify the situation, below is listed issues that one may have with the term "clone". In particular, the list details aspects or issues that would be important in distinguishing one concept or definition from another.

- Distinguishing merely similar to verbatim copy (cf. Dolly).
- Indicating the degree of similarity (what is a "*near*-clone"?).
- Indicating the category or type of difference [6].
- Indicating the provenance of the code (cf. copy-pasting, see copyright law or plagiarism rules).
- Distinguishing similar parts within a system (self-similarities) and between systems (plagiarism, copying, etc.).
- Distinguishing between helpful or desirable self-similarities or redundancies, and those that are detrimental or unwanted.
- Distinguishing between similarities one is interested in and similarities that one is not
- Indicating whether a clone is (easily) refactorable, or should be refactored.

The nature of concept naming makes it seems likely that each of the above aspects or distinctions will create a context in which either a new word or distinguishing modifier will be desired. For example, regarding the third point above, perhaps one could introduce the terms "intentional clone" versus "accidental clone"; the modifiers "intentional" and "accidental" would be used to indicate the distinction between copy-pasted and independently constructed "clones".

3 Review of Related English Definitions

Given the aspects and distinctions from the previous section, we may now turn to the question of what kind of terminology would be suitable for naming the various concepts. Here we list some common, related English terms. The motivation is to try to see if there are existing uses that closely match some of the concepts existing in our field. The source of the definitions are listed as a link, or using the following shorthands: OED = Oxford English Dictionary (online edition); Britannica = Encyclopedia Britannica (online edition); Wikipedia = Wikipedia (http://wikipedia.org). After introducing these, they are discussed with respect to what sort of definitions they may match or fail to match.

3.1 Definitions

Clone (noun)

- Any group of cells or organisms produced asexually from a single sexually produced ancestor. (OED)
- A thing produced in imitation of, or closely resembling, another; spec. a microcomputer designed to simulate the functions of another (usu. more expensive) model. (OED)
- Population of genetically identical cells or organisms that originated from a single cell or organism by nonsexual methods. (Britannica)
- An exact digital copy, indistinguishable from the original (http://www.wgcu.org/watch/hdtv_glossaryofterms.html)
- A clone is a computer system (both hardware and software) based on another company's system and designed to be compatible with it. (Wikipedia)
- Knockoff: an unauthorized copy or imitation (wordnet.princeton.edu)

Duplicate (noun)

- One of two things exactly alike, so that each is the 'double' of the other; especially, that which is made from or after the other. (OED)
- A second copy of a letter or official document, having the legal force of the original: whether made along with it, for separate custody or transmission, or prepared subsequently to take the place of the other in case of loss. (OED)
- A copy that corresponds to an original exactly; "he made a duplicate for the files" (wordnet.princeton.edu)

Redundant (adjective)

- Superabundant, superfluous, excessive. (OED)
- Of a language: containing material which is predictable from context or a knowledge of its structure; also of a language feature, predictable in this way. (OED)
- Repeated or duplicated unnecessarily. (aspin.asu.edu/geneinfo/glos-r.htm)
- Redundant describes computer or network system components, such as fans, hard disk drives, servers, operating systems, switches, and telecommunication links that are installed to back up primary resources in case they fail. (www.voip-architecture.com/glossary/glossary.html)
- Redundancy in information theory is the number of bits used to transmit a message minus the number of bits of actual information in the message. (Wikipedia)

Similar (adjective)

- Marked by correspondence or resemblance; "similar food at similar prices"; "problems similar to mine"; "they wore similar coats" (wordnet.princeton.edu)

- Of the same substance or structure throughout; homogeneous. (OED)
- Having a marked resemblance or likeness; of a like nature or kind. (OED)

Copy (noun)

- A writing transcribed from, and reproducing the contents of, another; a transcript. (OED)
- Something made or formed, or regarded as made or formed, in imitation of something else; a reproduction, image, or imitation. (OED)
- Imitate: reproduce someone's behavior or looks; "The mime imitated the passers-by"; "Children often copy their parents or older siblings" (wordnet.princeton.edu)
- Replicate: reproduce or make an exact copy of; "replicate the cell"; "copy the genetic information" (wordnet.princeton.edu)
- A reproduction or imitation of an original. (www.huntington.org/Education/lessons/SG-vocab1.htm)

Replica (noun)

- A copy, duplicate, or reproduction of a work of art; properly, one made by the original artist. (OED)
- An exact duplicate of the original, using the same materials and manufacturing techniques as were used to produce the original article. (www.aeroflight.co.uk/definitions.htm)
- Copy that is not the original; something that has been copied (wordnet.princeton.edu/perl/webwn)

Repetitious (adjective)

- Characterized or marked by repetition; especially: tediously repeating (Britannica)
- Abounding in, or characterized by, repetition, esp. of a tedious kind; tiresomely iterative. (OED)

Note that all but "similar" appear to relate primarily to things that are either exactly similar or very closely similar. At the time of this writing, we can think of no English word that means "quite similar but different in some way". The closest terms appear to come from copyright law: "derivative", "translation", and "adaptation".

3.2 Discussion

After reading through the definitions from the previous section, it appears that the English language may provide relatively little help in narrowing down choices, as definitions can be in conflict, or may be too general. Here we discuss the definitions above in reference to how they might be used:

Clone. In the biological context, a clone by definition requires copying (asexual reproduction), however in the retail case a clone can be constructed completely independently so long as it appears to be identical (or perhaps very nearly similar) to the original (games, designer purses, etc.). Thus, depending upon the metaphor used, we could consider clones as verbatim copy-pastes, or merely identical or very nearly identical pieces of software without any specific reference to provenance. These, essentially, are opposites. One wonders if the term should simply be avoided because of these conflicting interpretations. While the phrase "free as in freedom, not free as in beer" has been used elsewhere to distinguish two different connotations of the term "free", it seems like a poor solution to use a analogous phrases such as "clone as in Dolly, not clone as in $10 Rolex knockoff."

Duplicate. By its usage it appears to correspond closely to the notion of a "clone" created via copying. "Duplicate" seems generally to imply verbatim similarity.

Redundant. The term generally implies duplication that can be done without, except in the case of "redundant" computing infrastructure, which are not really redundant in the normal sense because otherwise there would be no point to the duplicated resources. Thus, like "clone", "redundant" seems to be a tainted word. Still, if anyone said that their software contains redundancies, it seems that the best strategy is to presume they could squeeze out some of the redundancies rather than thinking the redundant code is being used to ensure the system is robust. None of the other terms—perhaps save for "repetitious"—have the direct implication that the similar code is something that could be refactored to make it less redundant.

Copy. Appears to be the definition of "clone" requiring copying. Seems to be a better choice than clone for that purpose. Dolly = copy.

Replica. Appears to share the same connotations and overloaded meanings as "clone", however it more strongly connotes a closely similar object with *independent* derivation.

Repetitious. The quality of "redundant" code.

4 Suggestions and Conclusions

No single term appears to cover all previous connotations of "clone". Moreover, there appears to be voids in the terminology space. For example, the author does

not know of a term that precisely means "a snippet of code that is reasonably similar to another piece of code but is not exactly similar to it, may not be derived from it, may not be easy to refactor, and may not be desirable to be refactored". Given the confusion in the English words, it seems likely that the field will have to either develop its own terminology or else be careful in redefining known terms. For instance, in the world of genetics, the term "clone" has a significantly different meaning than that used in counterfeit merchandise. In this review, the terminology from copyright law appears to be under-used; as it by necessity its terminology must consider differences in provenance and similarity.

The authors look forward to a day when these naming issues are resolved; perhaps this report will provide some help in reaching that goal.

Acknowledgments

The author wishes to thank all the participants and organizers of the DRASIS workshop; this writeup is the product of their hard work.

References

1. Walenstein, A., Koschke, R., Merlo, E.: Duplication, redundancy, and similarity in software: Summary of Dagstuhl seminar 06301. [7] ISSN 1682–4405.
2. Walenstein, A., Jyoti, N., Li, J., Yang, Y., Lakhotia, A.: Problems creating task-relevant clone detection reference data. In van Deursen, A., Stroulia, E., Storey, M.A.D., eds.: WCRE, IEEE Computer Society (2003) 285–295
3. Koschke, R.: Survey of research on software clones. [7] ISSN 1682–4405.
4. Cordy, J.R.: Comprehending reality - practical barriers to industrial adoption of software maintenance automation. In: IWPC, IEEE Computer Society (2003) 196–206
5. Anderson, P., Frenzel, P., Koschke, R., Rieger, M.: Source code clone detectors: Overview and open problems. [7] ISSN 1682–4405.
6. Kapser, C., Godfrey, M.W.: Toward a taxonomy of clones in source code: A case study. In: Evolution of Large-scale Industrial Software Applications (ELISA), Amsterdam, The Netherlands (2003)
7. Proceedings of Dagstuhl Seminar 06301: Duplication, Redundancy, and Similarity in Software, Dagstuhl, Germany, Dagstuhl (2006) ISSN 1682–4405.