Sheep artwork taken from: https://iwsc2020.github.io/

Code Duplication on Stack Overflow

ICSE 2020 NIER

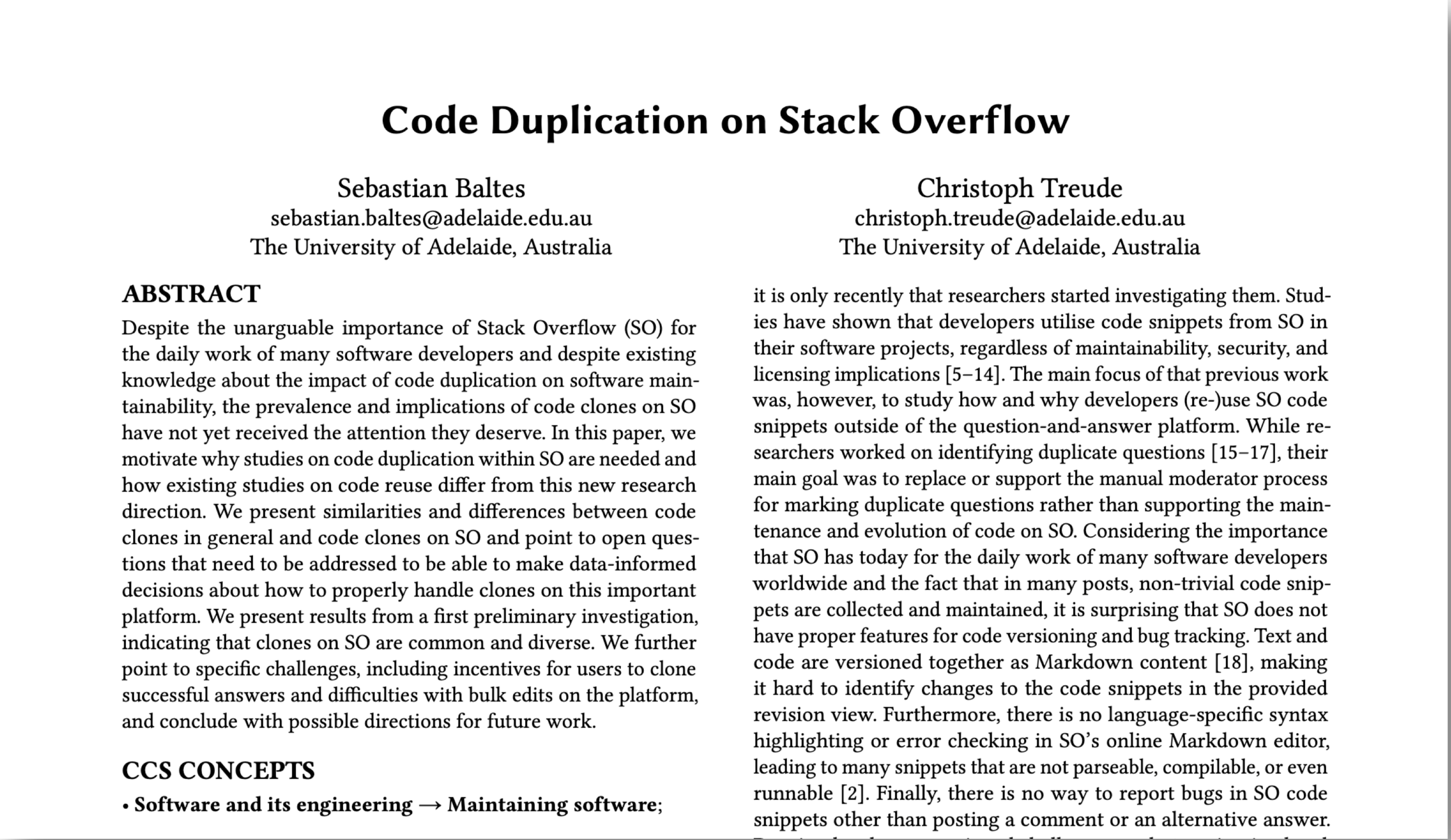
Sebastian Baltes

@s_baltes

empirical-software.engineering

THE UNIVERSITY of ADELAIDE

Thanks to my co-author!



Code Duplication on Stack Overflow

Sebastian Baltes
sebastian.baltes@adelaide.edu.au
The University of Adelaide, Australia

Christoph Treude
christoph.treude@adelaide.edu.au
The University of Adelaide, Australia

ABSTRACT

Despite the unarguable importance of Stack Overflow (SO) for the daily work of many software developers and despite existing knowledge about the impact of code duplication on software maintainability, the prevalence and implications of code clones on SO have not yet received the attention they deserve. In this paper, we motivate why studies on code duplication within SO are needed and how existing studies on code reuse differ from this new research direction. We present similarities and differences between code clones in general and code clones on SO and point to open questions that need to be addressed to be able to make data-informed decisions about how to properly handle clones on this important platform. We present results from a first preliminary investigation, indicating that clones on SO are common and diverse. We further point to specific challenges, including incentives for users to clone successful answers and difficulties with bulk edits on the platform, and conclude with possible directions for future work.

CCS CONCEPTS

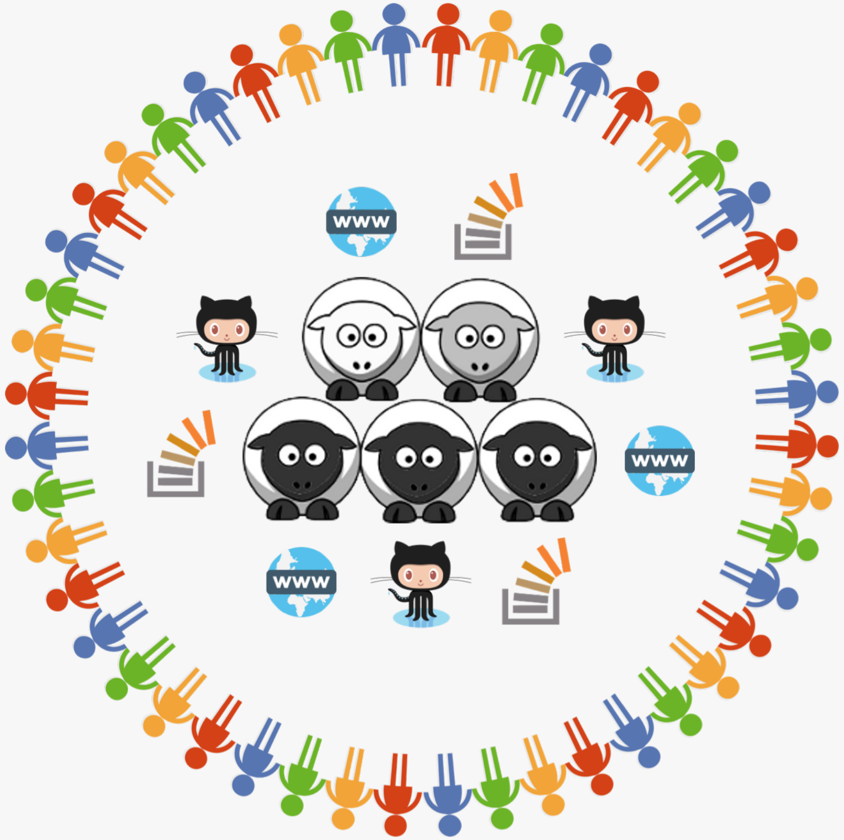
• **Software and its engineering → Maintaining software**;

it is only recently that researchers started investigating them. Studies have shown that developers utilise code snippets from SO in their software projects, regardless of maintainability, security, and licensing implications [5–14]. The main focus of that previous work was, however, to study how and why developers (re-)use SO code snippets outside of the question-and-answer platform. While researchers worked on identifying duplicate questions [15–17], their main goal was to replace or support the manual moderator process for marking duplicate questions rather than supporting the maintenance and evolution of code on SO. Considering the importance that SO has today for the daily work of many software developers worldwide and the fact that in many posts, non-trivial code snippets are collected and maintained, it is surprising that SO does not have proper features for code versioning and bug tracking. Text and code are versioned together as Markdown content [18], making it hard to identify changes to the code snippets in the provided revision view. Furthermore, there is no language-specific syntax highlighting or error checking in SO's online Markdown editor, leading to many snippets that are not parseable, compilable, or even runnable [2]. Finally, there is no way to report bugs in SO code snippets other than posting a comment or an alternative answer.

Yet another study about Stack Overflow?

Interestingly, there is one important aspect that previous research largely ignored...

Code Duplication on Stack Overflow from a Community Perspective

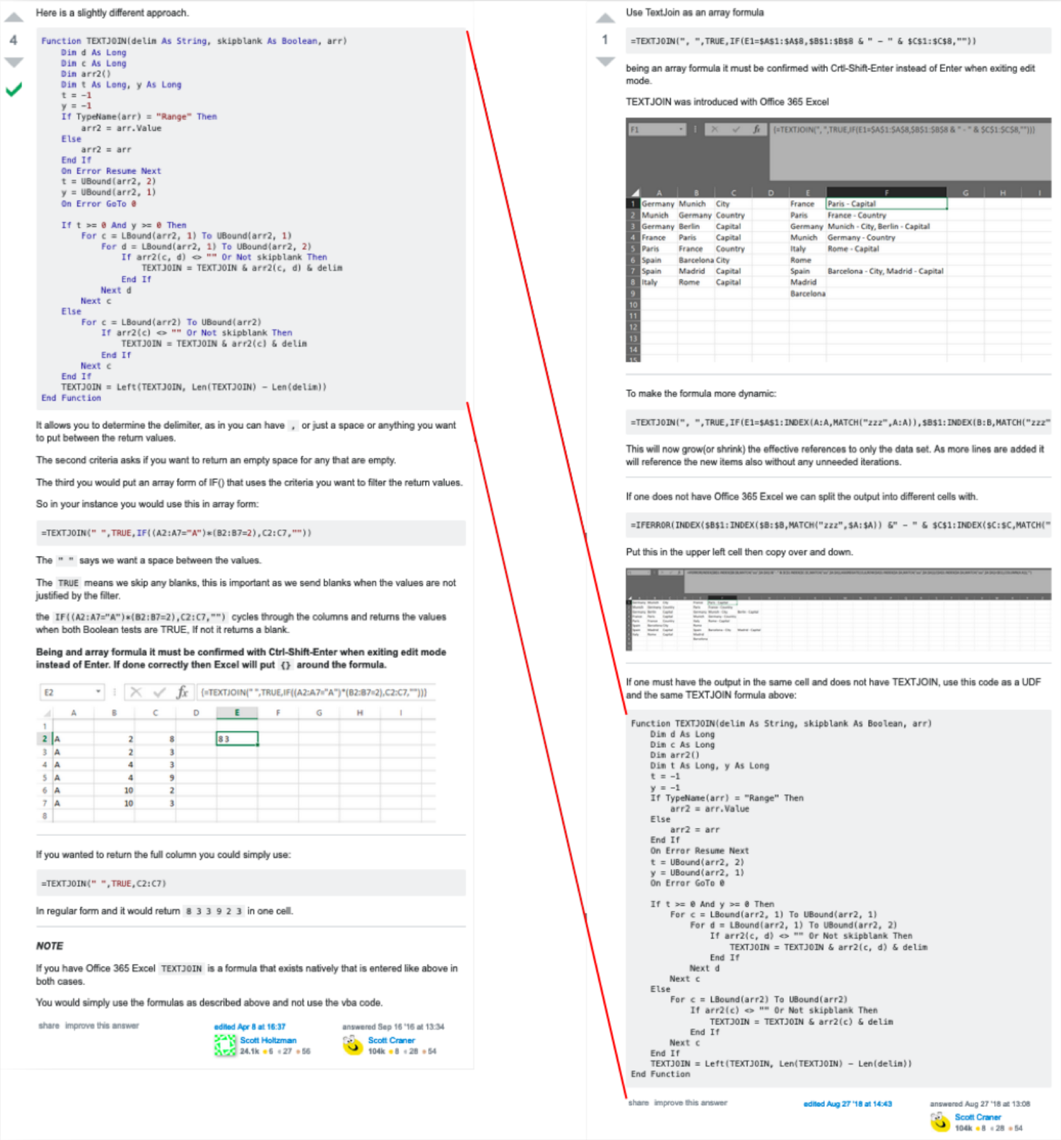


Scenario 1: Using clones to increase reputation

I could **re-post** a rather successful snippet wherever it fits **without referencing** the original answer to **accumulate** views and upvotes 🤔

Scenario 1: Using clones to increase reputation

- First usage of snippet in September 2016
- Overall **31 copies** of the same snippet, almost exclusively posted by the **same user**
- Now imagine someone finds a **bug**...



Here is a slightly different approach.

4

```
Function TEXTJOIN(delim As String, skipblank As Boolean, arr)
    Dim d As Long
    Dim c As Long
    Dim arr2()
    Dim t As Long, y As Long
    t = -1
    y = -1
    If TypeName(arr) = "Range" Then
        arr2 = arr.Value
    Else
        arr2 = arr
    End If
    On Error Resume Next
    t = UBound(arr2, 2)
    y = UBound(arr2, 1)
    On Error GoTo 0

    If t >= 0 And y >= 0 Then
        For c = LBound(arr2, 1) To UBound(arr2, 1)
            For d = LBound(arr2, 1) To UBound(arr2, 2)
                If arr2(c, d) <> "" Or Not skipblank Then
                    TEXTJOIN = TEXTJOIN & arr2(c, d) & delim
                End If
            Next d
        Next c
    Else
        For c = LBound(arr2) To UBound(arr2)
            If arr2(c) <> "" Or Not skipblank Then
                TEXTJOIN = TEXTJOIN & arr2(c) & delim
            End If
        Next c
    End If
    TEXTJOIN = Left(TEXTJOIN, Len(TEXTJOIN) - Len(delim))
End Function
```

It allows you to determine the delimiter, as in you can have `,` or just a space or anything you want to put between the return values.

The second criteria asks if you want to return an empty space for any that are empty.

The third you would put an array form of IF() that uses the criteria you want to filter the return values.

So in your instance you would use this in array form:

```
=TEXTJOIN(" ",TRUE,IF((A2:A7="A")*(B2:B7=2),C2:C7,""))
```

The `" "` says we want a space between the values.

The `TRUE` means we skip any blanks, this is important as we send blanks when the values are not justified by the filter.

the `IF((A2:A7="A")*(B2:B7=2),C2:C7,"")` cycles through the columns and returns the values when both Boolean tests are TRUE, If not it returns a blank.

Being and array formula it must be confirmed with Ctrl-Shift-Enter when exiting edit mode instead of Enter. If done correctly then Excel will put `{}` around the formula.

E2 {=TEXTJOIN(" ",TRUE,IF((A2:A7="A")*(B2:B7=2),C2:C7,""))}

	A	B	C	D	E	F	G	H	I
1									
2	A	2	8		8 3				
3	A	2	3						
4	A	4	3						
5	A	4	9						
6	A	10	2						
7	A	10	3						
8									

If you wanted to return the full column you could simply use:

```
=TEXTJOIN(" ",TRUE,C2:C7)
```

In regular form and it would return `8 3 3 9 2 3` in one cell.

NOTE

If you have Office 365 Excel `TEXTJOIN` is a formula that exists natively that is entered like above in both cases.

You would simply use the formulas as described above and not use the vba code.

share improve this answer

edited Apr 8 at 16:37
Scott Holtzman
24.1k ● 6 ● 27 ● 56

answered Sep 16 '16 at 13:34
Scott Craner
104k ● 8 ● 28 ● 54

Use TextJoin as an array formula

1

```
=TEXTJOIN(", ",TRUE,IF(E1=$A$1:$A$8,$B$1:$B$8 & " - " & $C$1:$C$8,""))
```

being an array formula it must be confirmed with Crtl-Shift-Enter instead of Enter when exiting edit mode.

TEXTJOIN was introduced with Office 365 Excel

F1 {=TEXTJOIN(", ",TRUE,IF(E1=$A$1:$A$8,$B$1:$B$8 & " - " & $C$1:$C$8,""))}

	A	B	C	D	E	F	G	H	I
1	Germany	Munich	City		France	Paris - Capital			
2	Munich	Germany	Country		Paris	France - Country			
3	Germany	Berlin	Capital		Germany	Munich - City, Berlin - Capital			
4	France	Paris	Capital		Munich	Germany - Country			
5	Paris	France	Country		Italy	Rome - Capital			
6	Spain	Barcelona	City		Rome				
7	Spain	Madrid	Capital		Spain	Barcelona - City, Madrid - Capital			
8	Italy	Rome	Capital		Madrid				
9					Barcelona				
10									
11									
12									
13									
14									

To make the formula more dynamic:

```
=TEXTJOIN(", ",TRUE,IF(E1=$A$1:INDEX(A:A,MATCH("zzz",A:A)),$B$1:INDEX(B:B,MATCH("zzz"
```

This will now grow(or shrink) the effective references to only the data set. As more lines are added it will reference the new items also without any unneeded iterations.

If one does not have Office 365 Excel we can split the output into different cells with.

```
=IFERROR(INDEX($B$1:INDEX($B:$B,MATCH("zzz",$A:$A)) &" - " & $C$1:INDEX($C:$C,MATCH("
```

Put this in the upper left cell then copy over and down.

If one must have the output in the same cell and does not have TEXTJOIN, use this code as a UDF and the same TEXTJOIN formula above:

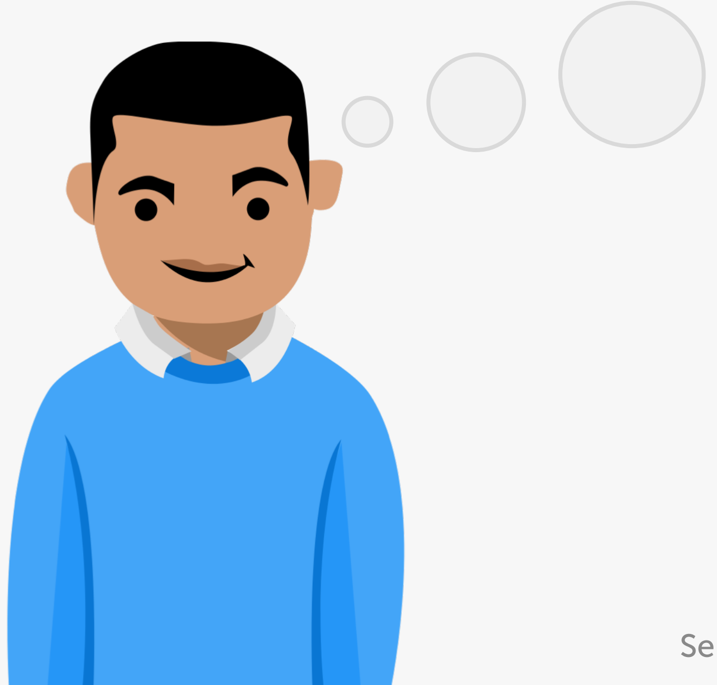
```
Function TEXTJOIN(delim As String, skipblank As Boolean, arr)
    Dim d As Long
    Dim c As Long
    Dim arr2()
    Dim t As Long, y As Long
    t = -1
    y = -1
    If TypeName(arr) = "Range" Then
        arr2 = arr.Value
    Else
        arr2 = arr
    End If
    On Error Resume Next
    t = UBound(arr2, 2)
    y = UBound(arr2, 1)
    On Error GoTo 0

    If t >= 0 And y >= 0 Then
        For c = LBound(arr2, 1) To UBound(arr2, 1)
            For d = LBound(arr2, 1) To UBound(arr2, 2)
                If arr2(c, d) <> "" Or Not skipblank Then
                    TEXTJOIN = TEXTJOIN & arr2(c, d) & delim
                End If
            Next d
        Next c
    Else
        For c = LBound(arr2) To UBound(arr2)
            If arr2(c) <> "" Or Not skipblank Then
                TEXTJOIN = TEXTJOIN & arr2(c) & delim
            End If
        Next c
    End If
    TEXTJOIN = Left(TEXTJOIN, Len(TEXTJOIN) - Len(delim))
End Function
```

share improve this answer

edited Aug 27 '18 at 14:43

answered Aug 27 '18 at 13:08
Scott Craner
104k ● 8 ● 28 ● 54

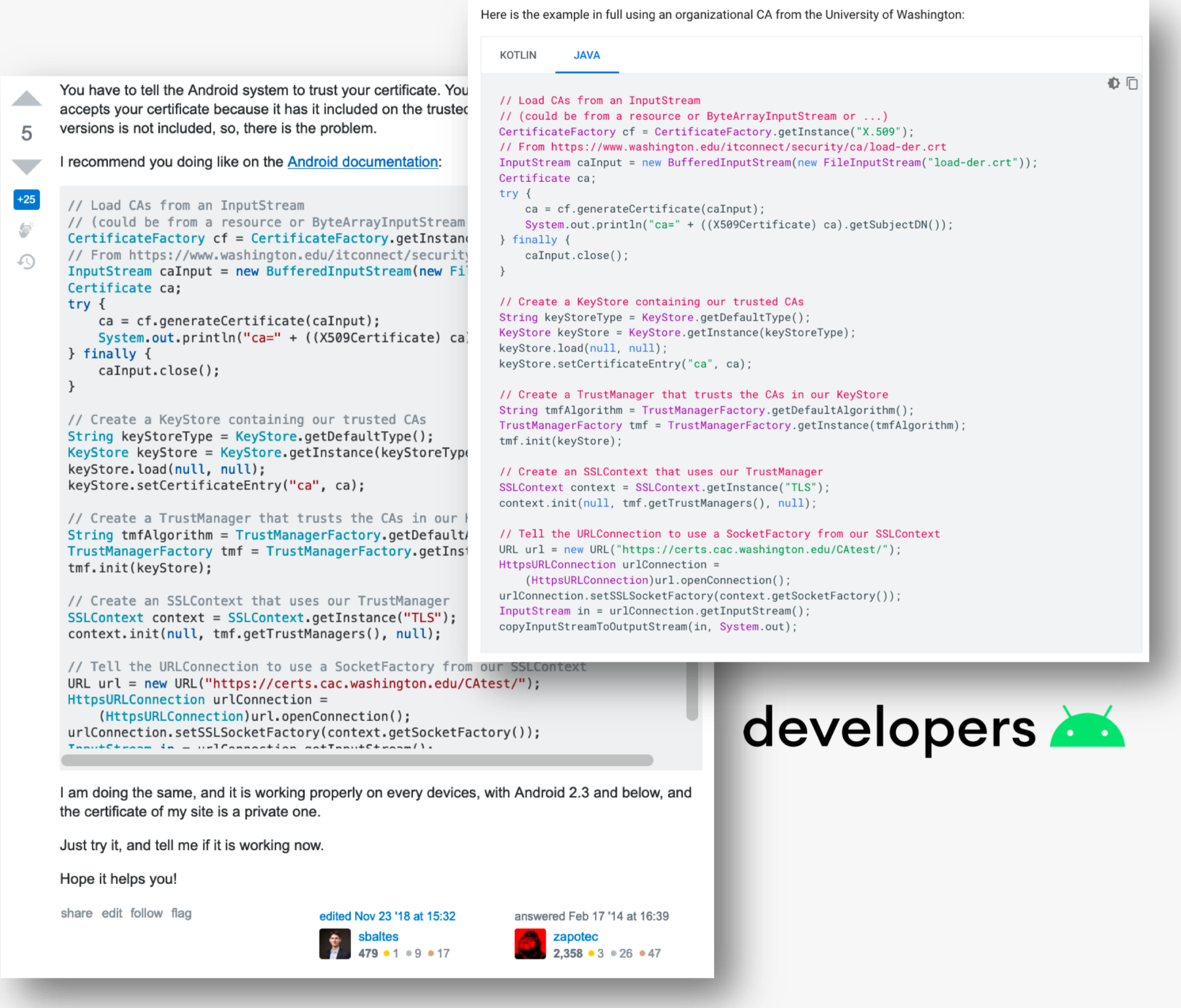
Scenario 2: Maintaining code copied into SO



To make it more likely that my post gets upvoted, I could **copy code** from **external** documentation resources into my answer instead of just referencing it 🤔

Scenario 2: Using clones to increase reputation

- Happens frequently
- Potential **licensing** issues
- What if the authoritative source gets **updated**?



5

+25

You have to tell the Android system to trust your certificate. You accepts your certificate because it has it included on the trusted versions is not included, so, there is the problem.

I recommend you doing like on the Android documentation:

```
// Load CAs from an InputStream
// (could be from a resource or ByteArrayInputStream
CertificateFactory cf = CertificateFactory.getInstanc
// From https://www.washington.edu/itconnect/security
InputStream caInput = new BufferedInputStream(new Fi
Certificate ca;
try {
    ca = cf.generateCertificate(caInput);
    System.out.println("ca=" + ((X509Certificate) ca)
} finally {
    caInput.close();
}

// Create a KeyStore containing our trusted CAs
String keyStoreType = KeyStore.getDefaultType();
KeyStore keyStore = KeyStore.getInstance(keyStoreType
keyStore.load(null, null);
keyStore.setCertificateEntry("ca", ca);

// Create a TrustManager that trusts the CAs in our K
String tmfAlgorithm = TrustManagerFactory.getDefaultA
TrustManagerFactory tmf = TrustManagerFactory.getInst
tmf.init(keyStore);

// Create an SSLContext that uses our TrustManager
SSLContext context = SSLContext.getInstance("TLS");
context.init(null, tmf.getTrustManagers(), null);

// Tell the URLConnection to use a SocketFactory from our SSLContext
URL url = new URL("https://certs.cac.washington.edu/CAtest/");
HttpsURLConnection urlConnection =
    (HttpsURLConnection)url.openConnection();
urlConnection.setSSLSocketFactory(context.getSocketFactory());
```

I am doing the same, and it is working properly on every devices, with Android 2.3 and below, and the certificate of my site is a private one.

Just try it, and tell me if it is working now.

Hope it helps you!

share edit follow flag

edited Nov 23 '18 at 15:32
sbaltes
479 ●1 ●9 ●17

answered Feb 17 '14 at 16:39
zapotec
2,358 ●3 ●26 ●47

Here is the example in full using an organizational CA from the University of Washington:

KOTLIN JAVA

```
// Load CAs from an InputStream
// (could be from a resource or ByteArrayInputStream or ...)
CertificateFactory cf = CertificateFactory.getInstance("X.509");
// From https://www.washington.edu/itconnect/security/ca/load-der.crt
InputStream caInput = new BufferedInputStream(new FileInputStream("load-der.crt"));
Certificate ca;
try {
    ca = cf.generateCertificate(caInput);
    System.out.println("ca=" + ((X509Certificate) ca).getSubjectDN());
} finally {
    caInput.close();
}

// Create a KeyStore containing our trusted CAs
String keyStoreType = KeyStore.getDefaultType();
KeyStore keyStore = KeyStore.getInstance(keyStoreType);
keyStore.load(null, null);
keyStore.setCertificateEntry("ca", ca);

// Create a TrustManager that trusts the CAs in our KeyStore
String tmfAlgorithm = TrustManagerFactory.getDefaultAlgorithm();
TrustManagerFactory tmf = TrustManagerFactory.getInstance(tmfAlgorithm);
tmf.init(keyStore);

// Create an SSLContext that uses our TrustManager
SSLContext context = SSLContext.getInstance("TLS");
context.init(null, tmf.getTrustManagers(), null);

// Tell the URLConnection to use a SocketFactory from our SSLContext
URL url = new URL("https://certs.cac.washington.edu/CAtest/");
HttpsURLConnection urlConnection =
    (HttpsURLConnection)url.openConnection();
urlConnection.setSSLSocketFactory(context.getSocketFactory());
InputStream in = urlConnection.getInputStream();
copyInputStreamToOutputStream(in, System.out);
```

Issues when addressing Scenarios 1 & 2

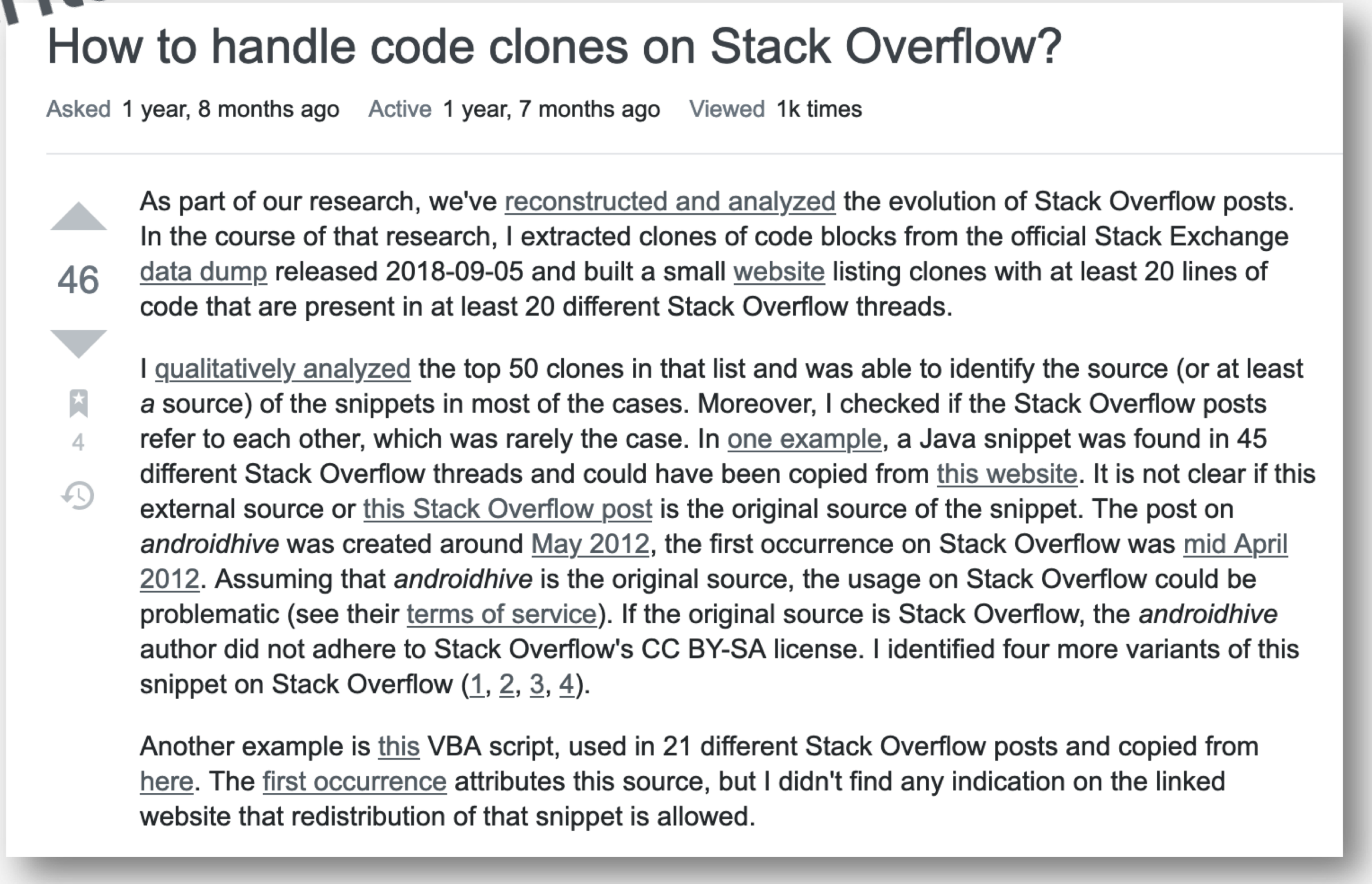
- Authors can **reject edits** linking the SO-internal clones

- SO's **rate limiting** prevents users from bulk editing posts
  (they would get reverted)

Community Involvement

stackoverflow META

https://meta.stackoverflow.com/q/375761/1974143



How to handle code clones on Stack Overflow?

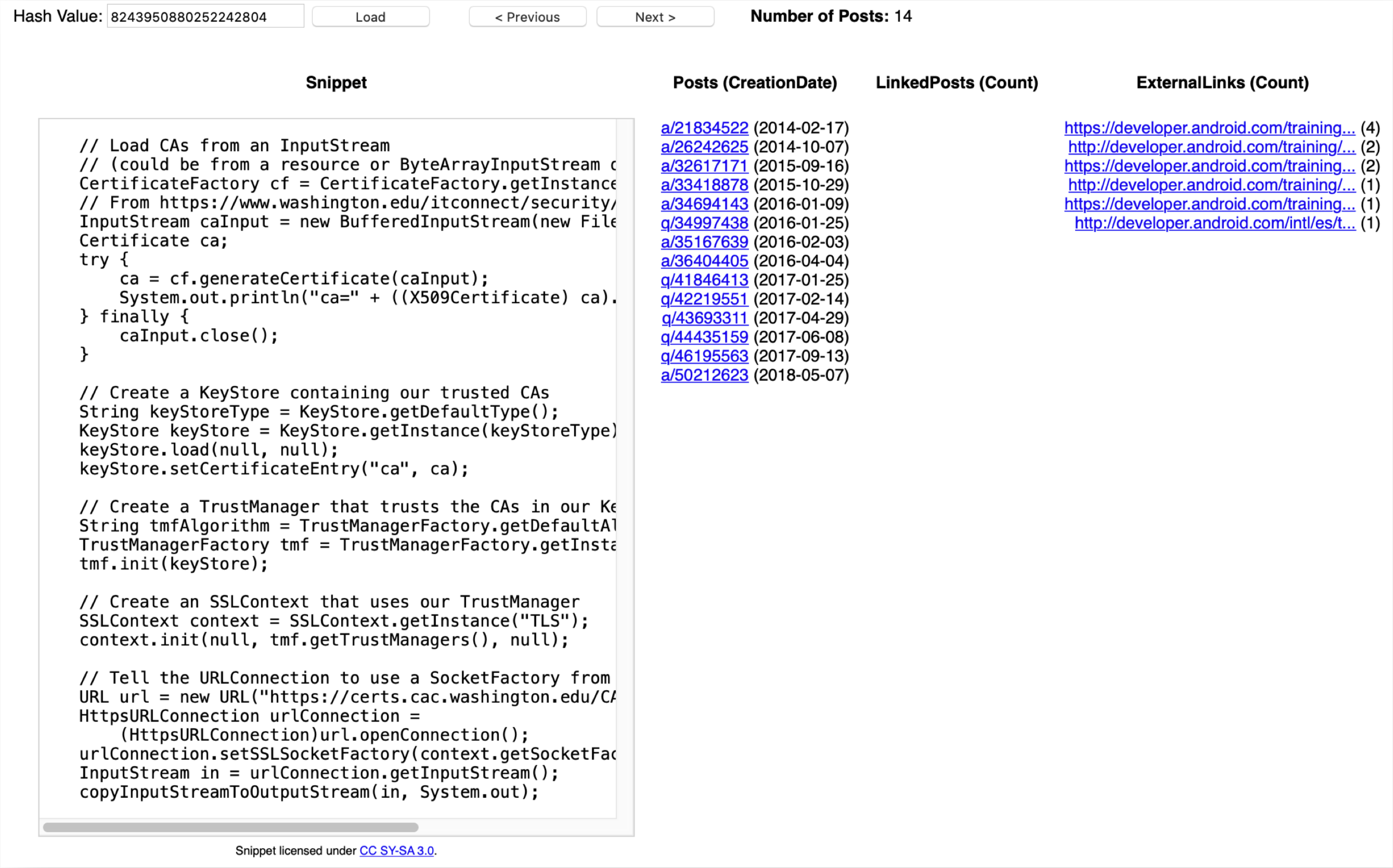
Asked 1 year, 8 months ago Active 1 year, 7 months ago Viewed 1k times

As part of our research, we've reconstructed and analyzed the evolution of Stack Overflow posts. In the course of that research, I extracted clones of code blocks from the official Stack Exchange data dump released 2018-09-05 and built a small website listing clones with at least 20 lines of code that are present in at least 20 different Stack Overflow threads.

I qualitatively analyzed the top 50 clones in that list and was able to identify the source (or at least *a* source) of the snippets in most of the cases. Moreover, I checked if the Stack Overflow posts refer to each other, which was rarely the case. In one example, a Java snippet was found in 45 different Stack Overflow threads and could have been copied from this website. It is not clear if this external source or this Stack Overflow post is the original source of the snippet. The post on *androidhive* was created around May 2012, the first occurrence on Stack Overflow was mid April 2012. Assuming that *androidhive* is the original source, the usage on Stack Overflow could be problematic (see their terms of service). If the original source is Stack Overflow, the *androidhive* author did not adhere to Stack Overflow's CC BY-SA license. I identified four more variants of this snippet on Stack Overflow (1, 2, 3, 4).

Another example is this VBA script, used in 21 different Stack Overflow posts and copied from here. The first occurrence attributes this source, but I didn't find any indication on the linked website that redistribution of that snippet is allowed.

(Very Basic) Tool Support



Number of Posts: 14

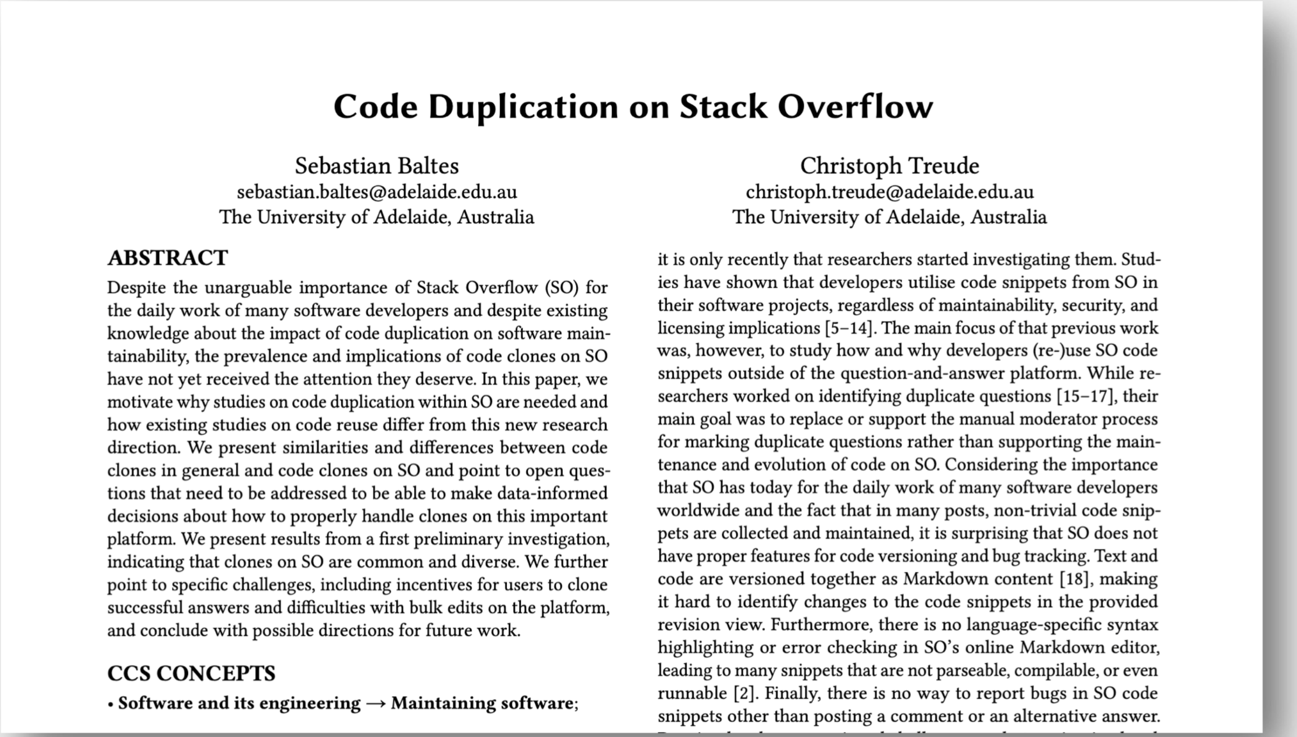
Snippet

```
// Load CAs from an InputStream
// (could be from a resource or ByteArrayInputStream c
CertificateFactory cf = CertificateFactory.getInstance
// From https://www.washington.edu/itconnect/security/
InputStream caInput = new BufferedInputStream(new File
Certificate ca;
try {
    ca = cf.generateCertificate(caInput);
    System.out.println("ca=" + ((X509Certificate) ca).
} finally {
    caInput.close();
}

// Create a KeyStore containing our trusted CAs
String keyStoreType = KeyStore.getDefaultType();
KeyStore keyStore = KeyStore.getInstance(keyStoreType)
keyStore.load(null, null);
keyStore.setCertificateEntry("ca", ca);

// Create a TrustManager that trusts the CAs in our Ke
String tmfAlgorithm = TrustManagerFactory.getDefaultAl
TrustManagerFactory tmf = TrustManagerFactory.getInsta
tmf.init(keyStore);

// Create an SSLContext that uses our TrustManager
SSLContext context = SSLContext.getInstance("TLS");
context.init(null, tmf.getTrustManagers(), null);

// Tell the URLConnection to use a SocketFactory from
URL url = new URL("https://certs.cac.washington.edu/CA
HttpsURLConnection urlConnection =
    (HttpsURLConnection)url.openConnection();
urlConnection.setSSLSocketFactory(context.getSocketFac
InputStream in = urlConnection.getInputStream();
copyInputStreamToOutputStream(in, System.out);
```

Snippet licensed under CC SY-SA 3.0.

Posts (CreationDate)	LinkedPosts (Count)	ExternalLinks (Count)
a/21834522 (2014-02-17)		https://developer.android.com/training... (4)
a/26242625 (2014-10-07)		http://developer.android.com/training/... (2)
a/32617171 (2015-09-16)		https://developer.android.com/training... (2)
a/33418878 (2015-10-29)		http://developer.android.com/training/... (1)
a/34694143 (2016-01-09)		https://developer.android.com/training... (1)
q/34997438 (2016-01-25)		http://developer.android.com/intl/es/t... (1)
a/35167639 (2016-02-03)		
a/36404405 (2016-04-04)		
q/41846413 (2017-01-25)		
q/42219551 (2017-02-14)		
q/43693311 (2017-04-29)		
q/44435159 (2017-06-08)		
q/46195563 (2017-09-13)		
a/50212623 (2018-05-07)		

http://research.sbaltes.com/so-clones/snippet-view.html?hashValue=8243950880252242804



Code Duplication on Stack Overflow

Sebastian Baltes
sebastian.baltes@adelaide.edu.au
The University of Adelaide, Australia

Christoph Treude
christoph.treude@adelaide.edu.au
The University of Adelaide, Australia

ABSTRACT

Despite the unarguable importance of Stack Overflow (SO) for the daily work of many software developers and despite existing knowledge about the impact of code duplication on software maintainability, the prevalence and implications of code clones on SO have not yet received the attention they deserve. In this paper, we motivate why studies on code duplication within SO are needed and how existing studies on code reuse differ from this new research direction. We present similarities and differences between code clones in general and code clones on SO and point to open questions that need to be addressed to be able to make data-informed decisions about how to properly handle clones on this important platform. We present results from a first preliminary investigation, indicating that clones on SO are common and diverse. We further point to specific challenges, including incentives for users to clone successful answers and difficulties with bulk edits on the platform, and conclude with possible directions for future work.

CCS CONCEPTS

• **Software and its engineering → Maintaining software**;

it is only recently that researchers started investigating them. Studies have shown that developers utilise code snippets from SO in their software projects, regardless of maintainability, security, and licensing implications [5–14]. The main focus of that previous work was, however, to study how and why developers (re-)use SO code snippets outside of the question-and-answer platform. While researchers worked on identifying duplicate questions [15–17], their main goal was to replace or support the manual moderator process for marking duplicate questions rather than supporting the maintenance and evolution of code on SO. Considering the importance that SO has today for the daily work of many software developers worldwide and the fact that in many posts, non-trivial code snippets are collected and maintained, it is surprising that SO does not have proper features for code versioning and bug tracking. Text and code are versioned together as Markdown content [18], making it hard to identify changes to the code snippets in the provided revision view. Furthermore, there is no language-specific syntax highlighting or error checking in SO's online Markdown editor, leading to many snippets that are not parseable, compilable, or even runnable [2]. Finally, there is no way to report bugs in SO code snippets other than posting a comment or an alternative answer.

Sebastian Baltes
@s_baltes

empirical-software.engineering