

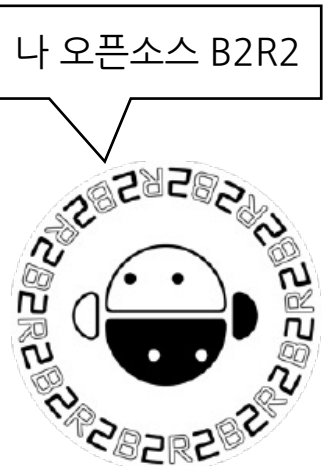
Binary Diff 및 Dataset 수집 도구 구현

[B2R2-BinDiff & Bianry-Gleaner]

팀36 윤형준 (산학 자유분반)
encrypt@kakao.com

I. 요약

- (1) binary diff 도구 B2R2-BinDiff 구현
- (2) binary diff를 위한 dataset을 수집하는 도구 Binary-Gleaner 구현
- B2R2-BinDiff:
 - 오픈 소스 바이너리 분석 플랫폼인 B2R2에 탑재
 - binary diff 기능
 - binary diff metric을 이용해 dataset을 평가할 수 있는 기능
- Binary-Gleaner:
 - binary diff에 필요한 파일 수준의 diff 정보와 binary-pair들로 이루어진 dataset을 수집하는 기능



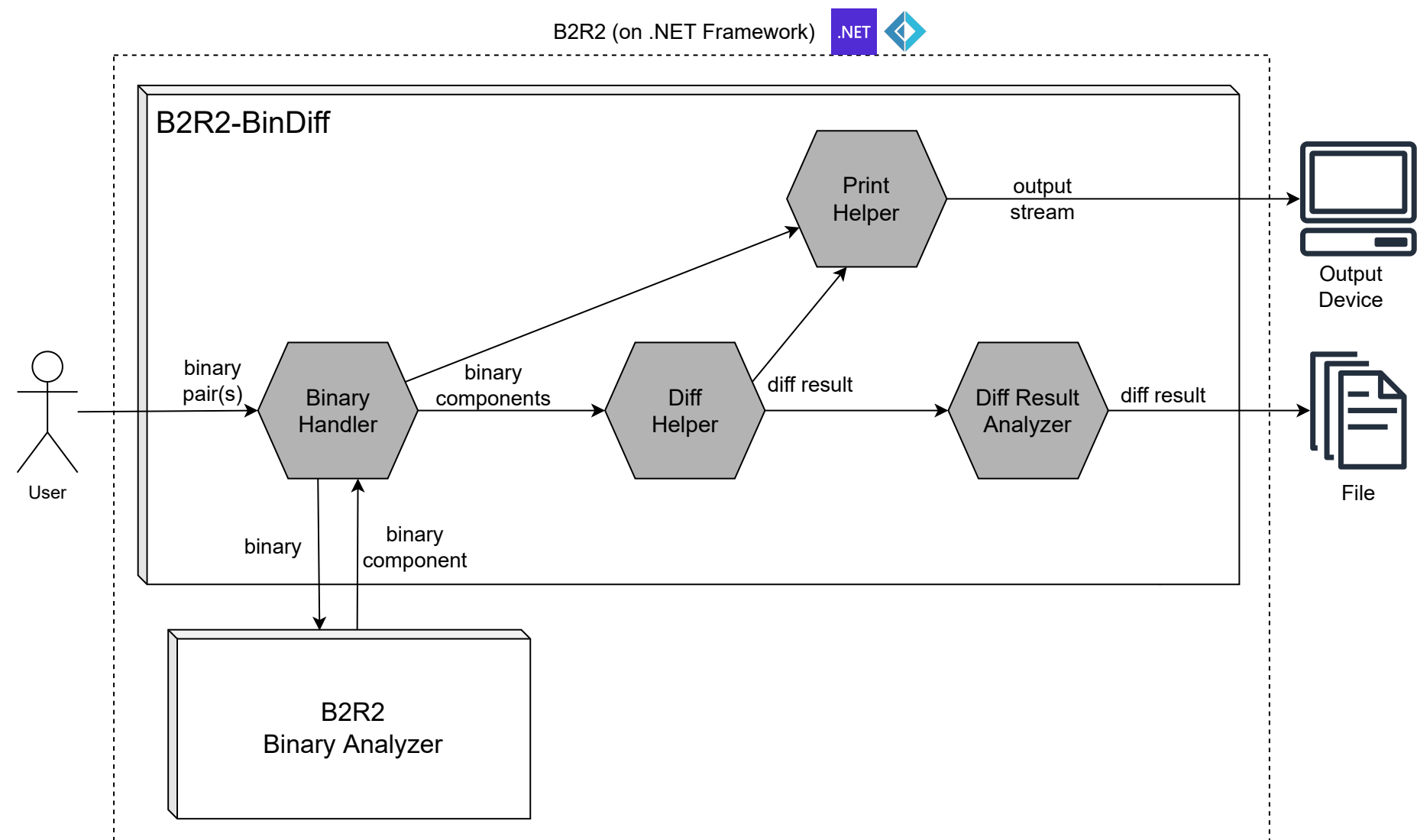
II. 배경

- binary란 실행 가능한 바이너리 파일(executable binary files)
- 바이너리 분석은 소프트웨어 보안을 실현하기 위해 필수적
- 상용 바이너리 분석 도구 IDA Pro는 모든 유료 라이선스 비용 약 1억 원
→ 무료 오픈 소스 바이너리 분석 기술 연구/개발이 필요한 실정
- binary diff:
 - 바이너리 분석 방법 중 하나
 - 동일한 바이너리의 서로 다른 두 버전의 차이를 식별하는 기술
- 현재 학계에는 binary diff 연구를 평가할 공통적인 양질의 dataset이 없다는 문제점이 존재
→ 오픈소스 binary diff 도구와 dataset 수집 도구를 만들자!



III. B2R2-BinDiff

시스템 구조도



주요 기능 소개

- 두 바이너리를 diff한 결과를 출력
 - CODE 영역은 disassembly하여 diff
 - DATA 영역은 Hex String을 diff
- binary-pair들로 이루어진 dataset을 metric으로 측정

```
<foo>
0000000000001169: endbr64
000000000000116d: push RBP
000000000000116e: mov RBP, RSP
0000000000001171: mov_dword_ptr [RBP-0x0], 0xa
0000000000001177: mov_dword_ptr [RBP-0x8], 0x14
0000000000001178: mov_dword_ptr [RBP-0xc], 0x14
000000000000117f: mov EAX, dword ptr [RBP-0xc]
0000000000001182: mov EAX, dword ptr [RBP-0x8]
0000000000001185: add EAX, 0x1e
0000000000001187: add EAX, 0x1e
000000000000118a: mov_dword_ptr [RBP-0x4], EAX
000000000000118d: mov EAX, dword ptr [RBP-0x4]
0000000000001190: pop RBP
0000000000001191: ret

<main>
0000000000001192: endbr64
0000000000001196: push RBP
0000000000001197: mov RBP, RSP
000000000000119a: lea RDI, qword ptr [0x2004]
00000000000011a1: call 0x141
00000000000011a6: mov EAX, 0x0
00000000000011ab: call 0x42
00000000000011b0: mov ESI, EAX
00000000000011b2: lea RDI, qword ptr [0x200c]
00000000000011b9: mov EAX, 0x0
00000000000011be: call 0x14e

00000000000011c3: mov EAX, 0x0
00000000000011c8: pop RBP
00000000000011c9: ret
00000000000011ca: nop word ptr [RAX+RAX+0x0]

<_libc_csu_init>
00000000000011d0: endbr64
00000000000011d4: push R15
00000000000011d6: lea R15, qword ptr [0x3db0]
00000000000011dd: push R14
00000000000011df: mov R14, RDX

<foo>
0000000000001169: endbr64
000000000000116d: push RBP
000000000000116e: mov RBP, RSP
0000000000001171: mov_dword_ptr [RBP-0x0], 0xa
0000000000001177: mov_dword_ptr [RBP-0x8], 0x14
0000000000001178: mov_dword_ptr [RBP-0xc], 0x14
000000000000117f: mov EAX, dword ptr [RBP-0xc]
0000000000001182: mov EAX, dword ptr [RBP-0x8]
0000000000001185: add EAX, 0x1e
0000000000001187: add EAX, 0x1e
000000000000118a: mov_dword_ptr [RBP-0x4], EAX
000000000000118d: mov EAX, dword ptr [RBP-0x4]
0000000000001190: pop RBP
0000000000001191: ret

<main>
0000000000001192: endbr64
0000000000001196: push RBP
0000000000001197: mov RBP, RSP
000000000000119a: lea RDI, qword ptr [0x2004]
0000000000001199: call 0x141
000000000000119e: mov EAX, 0x0
00000000000011a3: call 0x42
00000000000011a8: mov ESI, EAX
00000000000011aa: lea RDI, qword ptr [0x2009]
00000000000011b1: mov EAX, 0x0
00000000000011b6: call 0x14e
00000000000011b9: mov EAX, 0x0
00000000000011c0: pop RBP
00000000000011c1: ret
00000000000011c2: nop word ptr [RAX+RAX+0x0]
00000000000011c3: nop_dword_ptr [RAX+0x0]

<_libc_csu_init>
00000000000011d0: endbr64
00000000000011d4: push R15
00000000000011d6: lea R15, qword ptr [0x3db0]
00000000000011dd: push R14
00000000000011df: mov R14, RDX
```

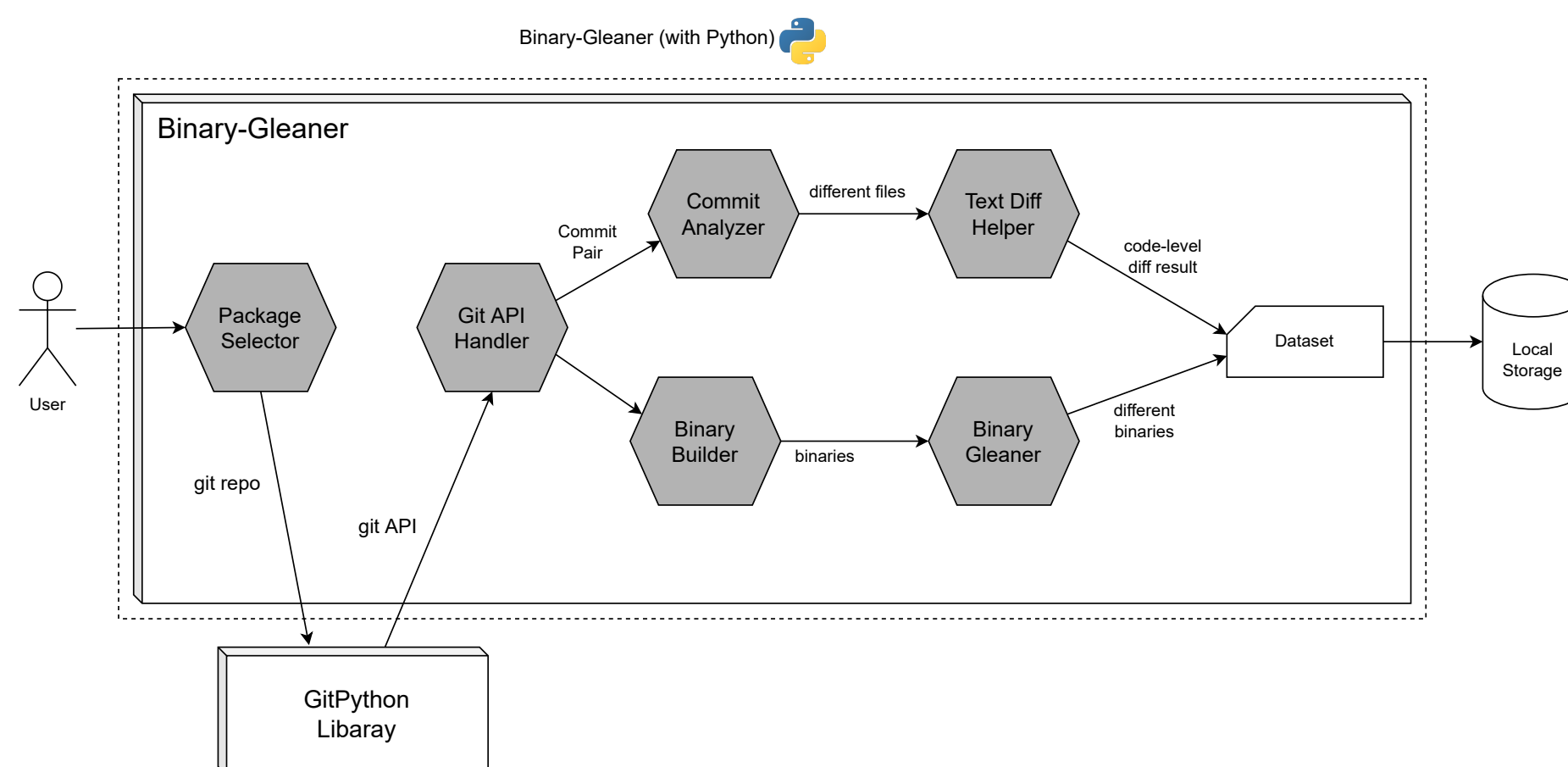
[binary diff 결과 (CODE 영역)]

기대효과

- Binary Diff 기능 제공
→ 바이너리 분석가 및 보안 전문가가 무료로 사용 가능
- Metric을 이용한 Binary Diff Dataset 측정 기능 제공
→ Binary Diff 관련 연구자가 연구에 활용 가능
- 모든 소스코드는 오픈 소스하여 공개
→ 오픈 소스 소프트웨어 및 보안 생태계에 기여

IV. Binary-Gleaner

시스템 구조도



주요 기능 소개

- binary diff에 필요한 dataset을 수집
 - 특정 repo(예: GNU binutils)의 모든 commit별 data 수집
 - 각 commit-pair의 high-level diff 결과 수집
 - 각 binary-pair 수집

```
File: info-20220520T204101-92bb0228c8293ec78c0efcd556b1f115b6e1b3f4

1 <?xml version='1.0' encoding='utf-8'?>
2 <Info>
3   <Commit>
4     <hexsha>92bb0228c8293ec78c0efcd556b1f115b6e1b3f4</hexsha>
5     <parent>acd0955bc118d14dd32c08fd8a6b2ca7fa4e294c</parent>
6     <committed_datetime>2022-05-20 20:41:01+01:00</committed_datetime>
7     <num_files>3</num_files>
8     <diff_files>
9       <file>gdb/break-catch-throw.c</file>
10      <file>gdb/breakpoint.c</file>
11      <file>gdb/breakpoint.h</file>
12    </diff_files>
13    <message />
14  </Commit>
15  <Files>
16    <File>
17      <path>gdb/break-catch-throw.c</path>
18      <type>M</type>
19      <stats>
20        <insertions>1</insertions>
21        <deletions>5</deletions>
22        <lines>6</lines>
23      </stats>
24      <curr_contents>/* Everything about catch/throw catchpoints, for GDB.
```

[binary diff 결과 (CODE 영역)]

기대효과

- Binary Diff Dataset 수집 기능 제공
→ 학계의 Dataset 부재 문제 해결
→ 다양한 이전 연구 및 앞으로의 binary diff 연구들을 체계적으로 평가 가능
- high-level data(소스코드 수준)과 low-level data(binary-pair) 동시 제공
→ 소스코드 수준의 diff와 binary 수준의 diff의 양상을 이용한 연구 가능
→ 오픈 소스 소프트웨어 및 보안 생태계에 기여