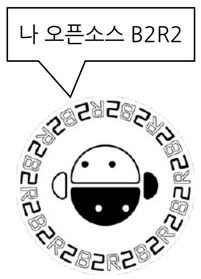
Binary Diff 및 Dataset 수집 도구 구현

[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



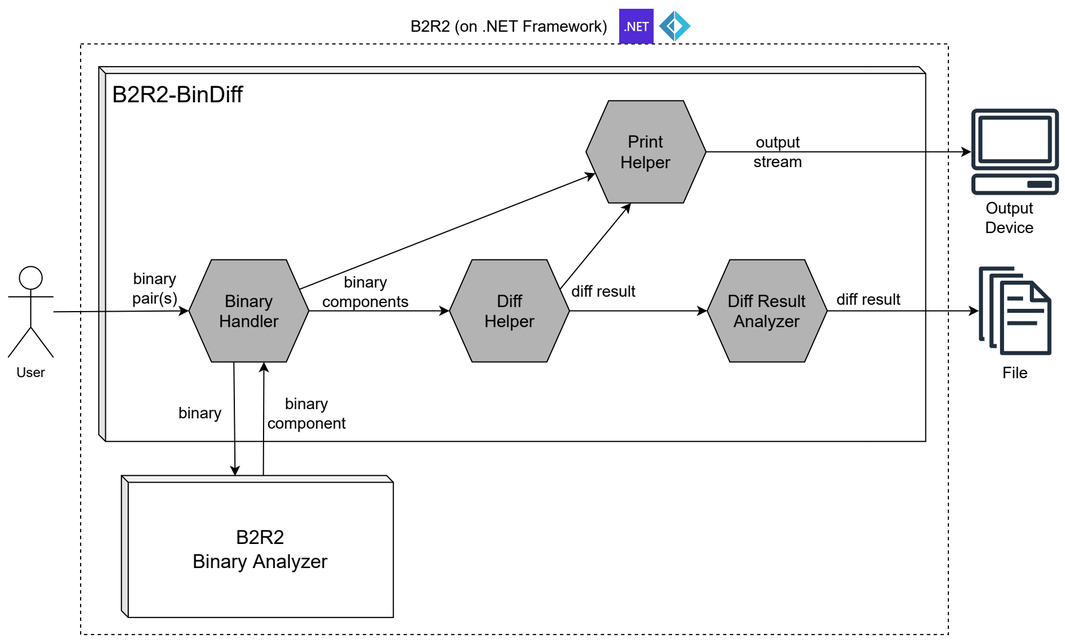
II. 배경

- binary란 실행 가능한 바이너리 파일(executable binary files)
- 바이너리 분석은 소프트웨어 보안을 실현하기 위해 필수적
- 상용 바이너리 분석 도구 IDA Pro는 모든 유료 라이센스 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!

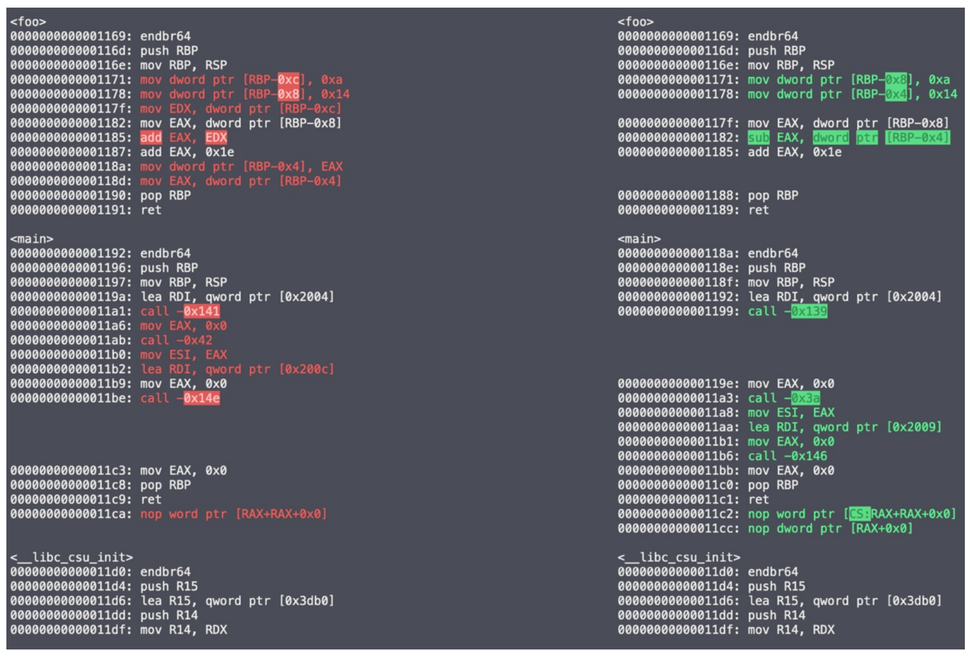
III. B2R2-BinDiff

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



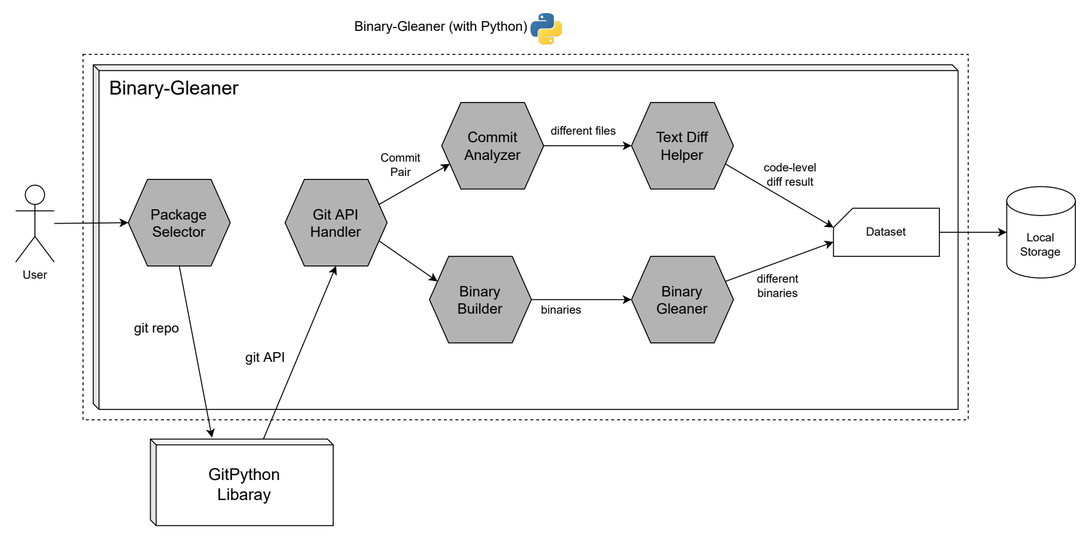
[binary diff 결과 (CODE 영역)]

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

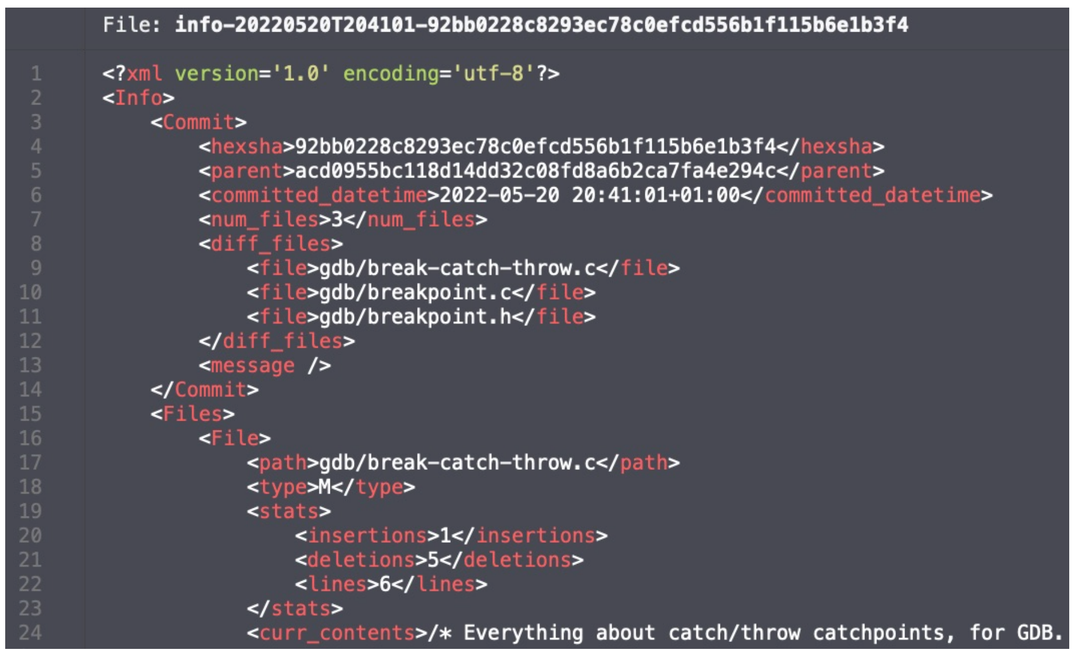
IV. Binary-Gleaner

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



[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