

# Mapping known vulnerabilities to SWH

Valentin Lorentz (@vlorentz)

Software Heritage

2025-10-16 - SWHSec Plenary



## Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

# Outline

- 1 osv.dev database
- 2 Versions in affected field
- 3 Limitations
- 4 Expanding `fixed` with cherry-picks
- 5 Recomputing introduced
- 6 Conclusion



# osv.dev database

- aggregates other databases, mostly in OSV format themselves
- each report contains a bunch of ids and a description
- affected software packages
  - 38k unique (1.6M total) are in SWH
  - 55k unique (1.4M) are in ecosystems not supported by SWH. Mostly RHEL (Red Hat) and Ubuntu Pro
  - 800 unique (56k) are missing from SWH
  - 7k unique are bogus

# Example of affected

```
{  "affected": [{
    "package": {
      "ecosystem": "PyPI",
      "name": "pyexample"
    },
    "ranges": [{
      "type": "GIT",
      "repo": "https://github.com/...",
      "events": [ [ ] ]
    }
  ]
}]
}
```

# Outline

- 1 osv.dev database
- 2 Versions in affected field
- 3 Limitations
- 4 Expanding `fixed` with cherry-picks
- 5 Recomputing introduced
- 6 Conclusion



# Versions in affected field

- either a range of version numbers, or of Git commits
- Four types of events:
  - `introduced`
  - `fixed`
  - `last_affected`
  - `limit`, which is like `last_affected` but restricts to a path from `introduced`

# Example of events

```
{  
  "events": [  
    { "introduced": "1.0" },  
    { "fixed": "1.2" },  
    { "introduced": "2.0" },  
    { "fixed": "2.1" },  
  ]  
}
```

# Outline

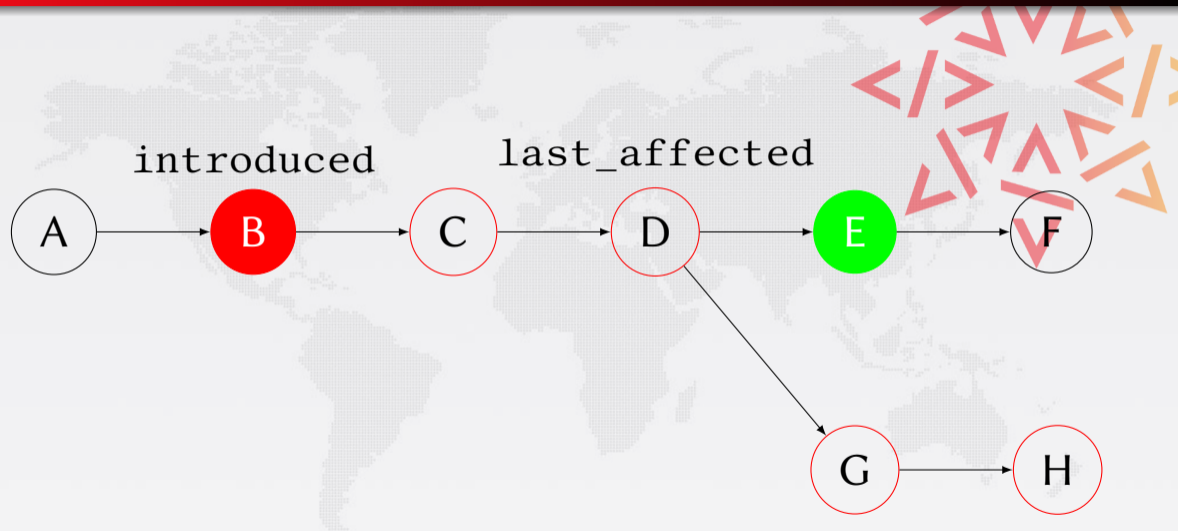
- 1 osv.dev database
- 2 Versions in affected field
- 3 Limitations**
- 4 Expanding *fixed* with cherry-picks
- 5 Recomputing *introduced*
- 6 Conclusion



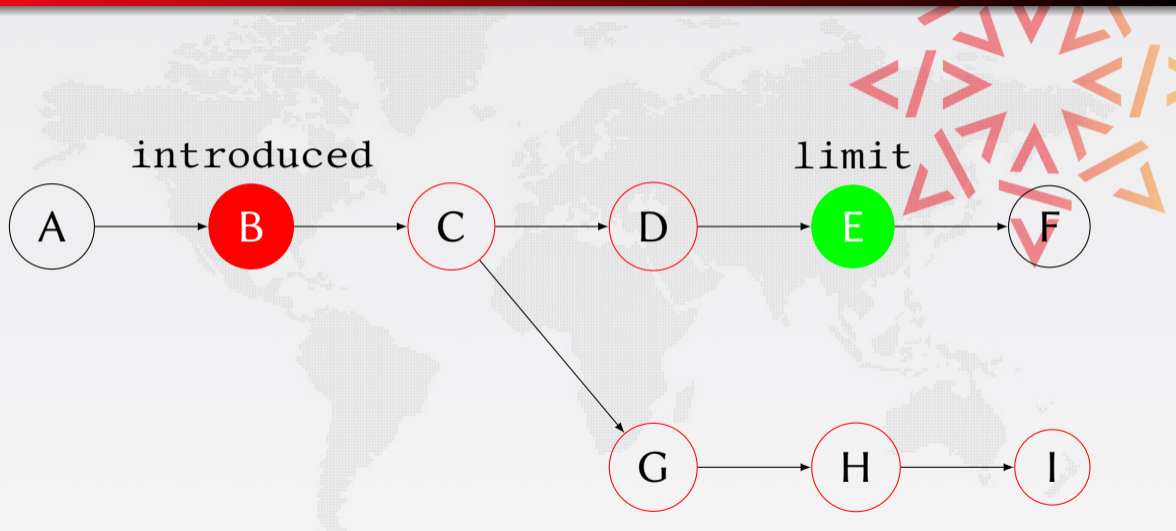
# Limitations

- `introduced` is often "0", meaning "assume it's vulnerable since the very first version"
- `fixed` is not exhaustive
- `last_affected` and `limit` are lies

# last\_affected



# limit

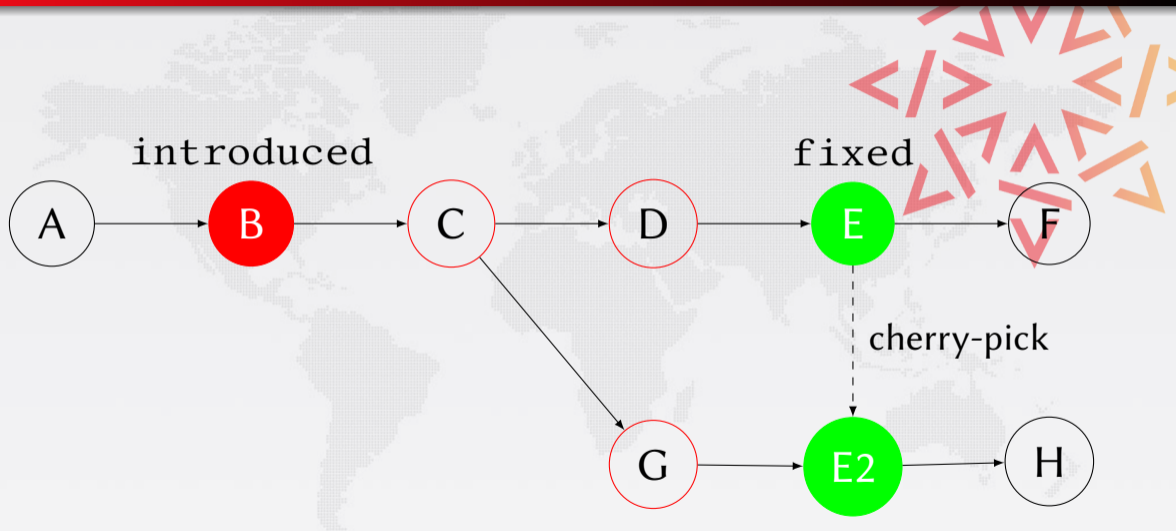


# Outline

- 1 osv.dev database
- 2 Versions in affected field
- 3 Limitations
- 4 Expanding fixed with cherry-picks
- 5 Recomputing introduced
- 6 Conclusion



# fixed



# Expanding `fixed` with cherry-picks

- It's not exhaustive because it's missing cherry-picks
- Cherry-pick detection based on commit message.
  - 287M are not deemed to be cherry-picks
  - 6.3M have at least one valid "cherry picked from commit" stanza (300k have at least two).  
Out of these:
    - 0.4M reference unknown commits and don't have a repository URL
    - 6.4M reference known commits

# Outline

- 1 osv.dev database
- 2 Versions in affected field
- 3 Limitations
- 4 Expanding `fixed` with cherry-picks
- 5 **Recomputing introduced**
- 6 Conclusion



# Recomputing introduced

- *SZZ* algorithm from *MSR2005*
- further refinements over the years
- *V-SZZ* in *ICSE2022*
- Implementation for *SWH*: work in progress, currently limited by download of all 600M relevant contents in a local RocksDB



# Outline

- 1 osv.dev database
- 2 Versions in affected field
- 3 Limitations
- 4 Expanding `fixed` with cherry-picks
- 5 Recomputing introduced
- 6 **Conclusion**



# Conclusion

- Matrix: `#swh-devel:matrix.org / @vlorentz:trix.re`
- Email: `vlorentz@softwareheritage.org`
- Work-in-progress code at <https://gitlab.softwareheritage.org/vlorentz/osv>