How we ~~managed~~ going to automate SBOM generation for a large legacy project

Thorsten Behrens
thorsten.behrens@collabora.com

FOSDEM

Collabora Online

Overview & Challenges

Goals

- have LibreOffice-lineage software ready for CRA times
- provide full dependency relations
- ultimately: complete transparency and traceability for all product artifacts

Challenges

- impressive mix of technology
  - JS/TS 3rd party modules
  - >100 native code dependencies
  - edge-case fonts & dictionaries
- no automated solutions
- evolving standards & unclear responsibilities

FOSDEM

Let’s get going!

First steps

- assess current state (… there’s nothing ...)
- read up & talk to experts & decide on SBOMs
  - roll dice & pick SPDX over CycloneDX
- realize there’s no pre-existing tools to extract all components for us
- roll up sleeves & collect Collabora Online components manually...

Collabora Online components

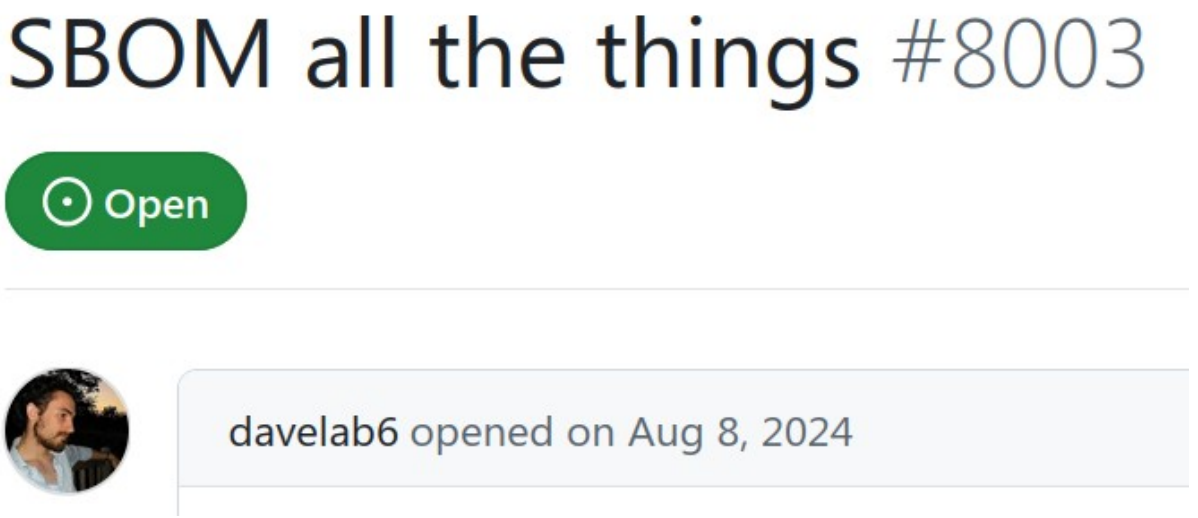
- looked around harder for automation, but ...
- collect JS deps manually
- notice we ship a font (for the admin console)
  - add another dependency manually
- review c/c++ dependencies
  - notice the generally terrible state – collect them manually
- minimally at least: update at least version info from build system!

Detour: Fonts & dictionaries

- Edge-case: fonts
  - they *do* have a license
  - and are binary artifacts
  - but are they relevant for CRA? well hinting tables are code…
- Edge-case: dictionaries
  - they *do* have a license
  - and are binary artifacts
  - and again, hyphenation patterns might be considered code...

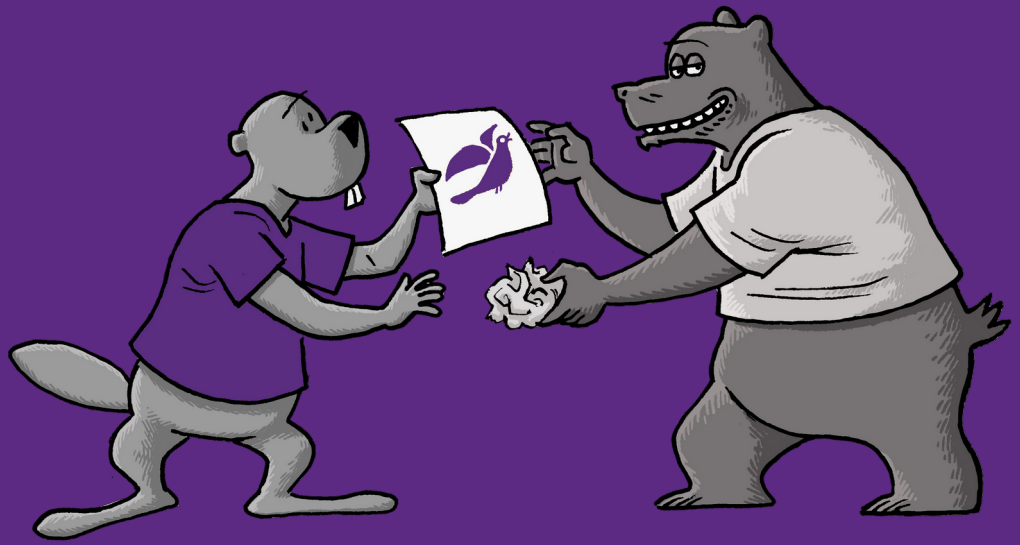
No luck with Fonts anyway

- https://github.com/google/fonts/issues/8003



FOSDEM

So what have we got?



Collabora Online SBOM

- commit from January 2025 :

```
SBOM: create and SPDX JSON file and install it with the package

Currently it contains only SBOM for the online part. Collabora
Office part will come in a follow-up commit.

Signed-off-by: Andras Timar <andras.timar@collabora.com>
Change-Id: Icfffd15e57ce58806a239e6beb4cc527f42750f0
```

main (#10988) · helm-collabora-online-1.1.55 ••• 24.04.12-mobile

Filter files...

- .gitignore
- Makefile.am
- cool-sbom-template.spdx.json
- scripts
  - create-sbom.py

4 files changed +447 -0 lines changed

.gitignore

```
@@ -183,3 +183,6 @@ autogen.input
183 183
184 184    # MacOS file manager metadata
185 185    .DS_Store
    186 +
    187 + # SBOM
    188 + collabora-online-sbom.spdx.json
```

Collabora Online SBOM

- `collabora-online-sbom.spdx.json`
  - at least lists all package content relationships – 26
  - version numbers coming from the build system
  - manually added `poco` statically-linked deps:
    - `libz, pcre, expat`
  - but: missing indirect outside-package dependencies
    - most prominently: `collaboraoffice-core`

FOSDEM

SBOMs for a GNU make project

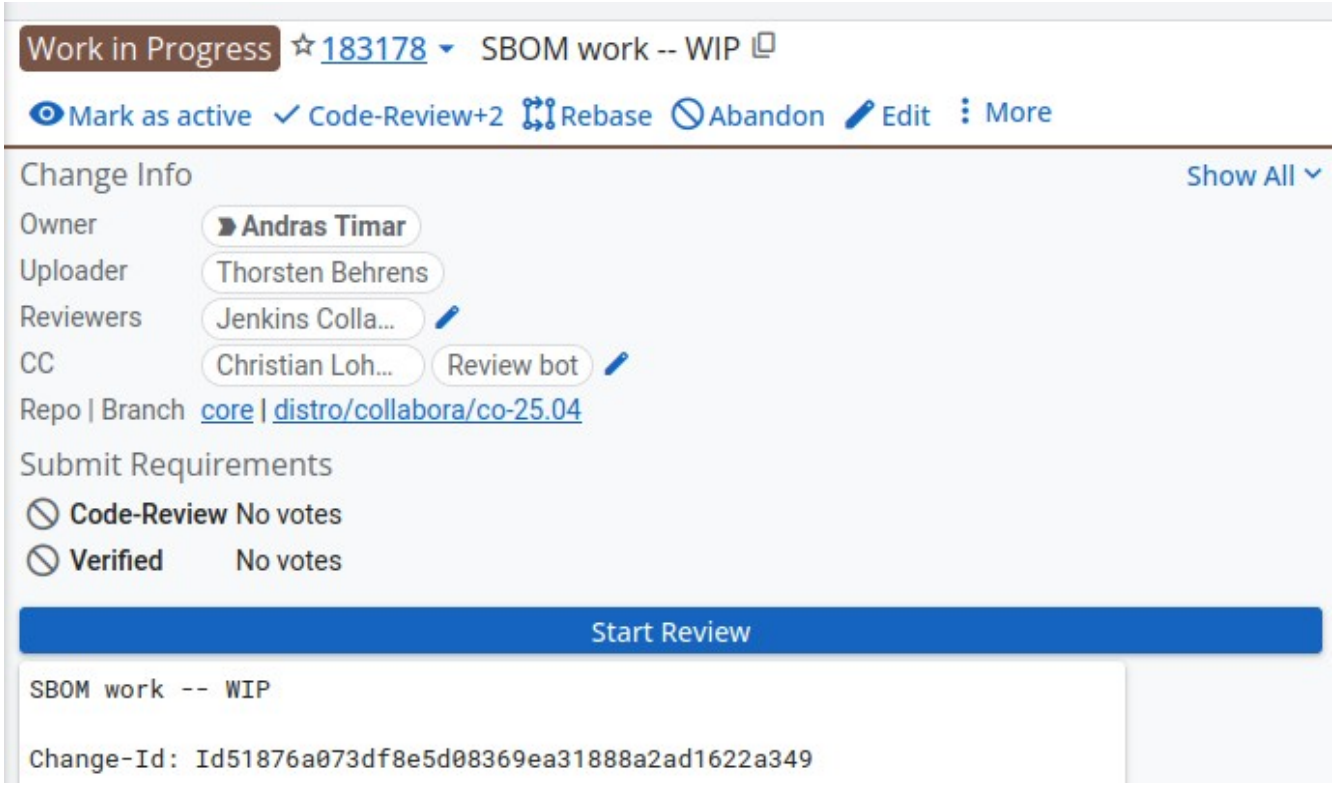
…and a very large one to boot.

LOKit core SBOM

- LibreOffice kit (core / collaboraoffice* packages) are complex
  - >100 3rd party libs
  - >50 fonts
  - >50 dictionaries
- build system knows about all of those
  - but again no existing, automated tooling available
  - clearly c/c++ projects & bespoke build systems are a challenge
- First cut: let’s linearly list all included dependencies that we ship!

LOKit core SBOM generation

- work-in-progress SBOM generation



LOKit core open issues

- for vulnerability assessment, we require all binary artifacts!
  - on disk, not just the packages...
  - including their dependency tree & linkage!
- so that needs:
  - SPDX 3.0.1 (according to BSI TR-03183)?
  - or perhaps better CycloneDX?
- we would *massively* benefit from our dependencies to provide SBOMs already

LOKit core todos

- ∀ “executables”
  - … including interpreted scripts *and* shared libraries
  - must be described as a *component* in an SBOM
  - system libraries that are linked against must be identified
    - but luckily not described
- ∀ installed components - i.e. executables, libraries & archives
  - we need:
    - name
    - SHA512 checksum
    - executable/archive/container flags

LOKit core todos (cont.)

- Information required for each bundled *external* component:
  - canonical name
  - version
  - packageURL (“generic” typically)
  - cpe23 identifier if it exists – these are centrally registered
  - declared license
  - concluded license
  - Source-URL
  - Source-hash
  - ***!!files in the installed package!!***

FOSDEM

“please why does libXYZ not have an SBOM yet?”

…and why oh why must we be at the top of the stack?

Well duh...

- ***all files in the installation package ?!***
  - how can we derive (automatically), which files originate from which external?
  - how can we derive (automatically) dynamic linking dependencies?
  - How can we derive (automatically) static linking dependencies?
- ***sadly for c/c++, there is no existing tool that could read some sensible package management metadata - we need a fully custom generator***

An actual plan

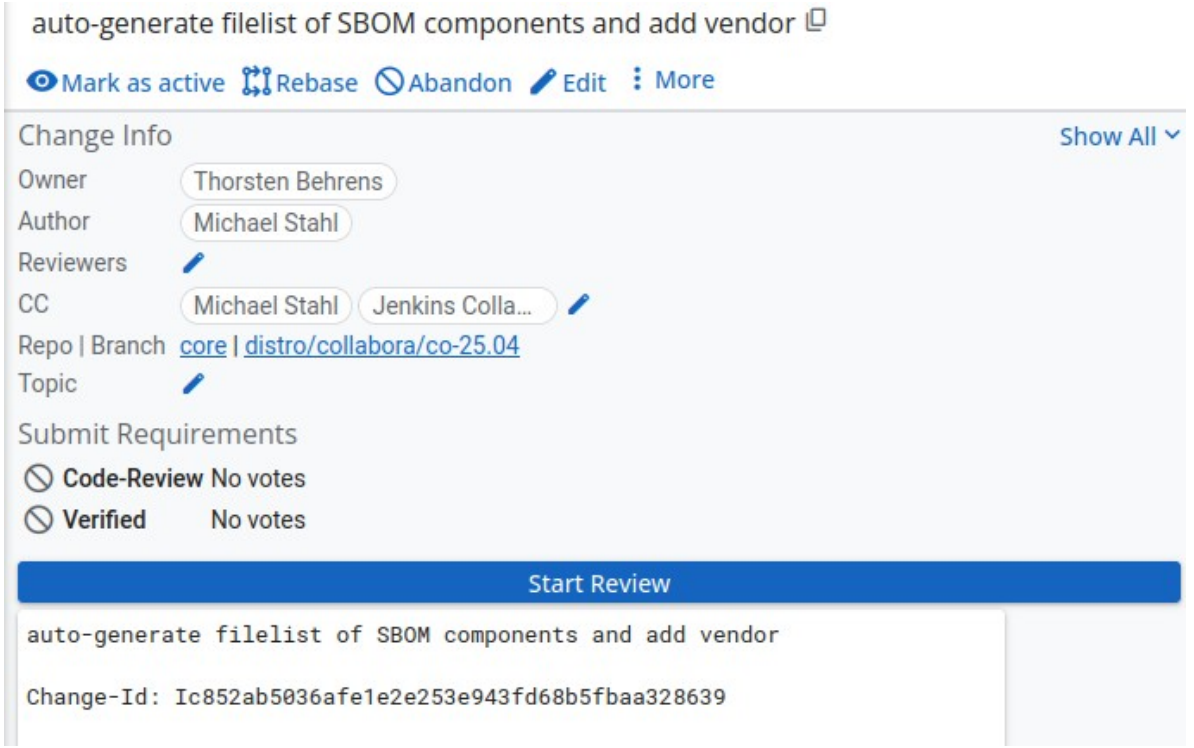
- bonus challenge
  - should also work on macOS and Windows and iOS and Android (where we ship)
- therefore: SBOMs must be generated during the build
  - otherwise *infeasible* to manually maintain
  - as they *will* depend on *current* build configuration
- the good news: technically, the build system knows all of this!

An actual plan (cont.)

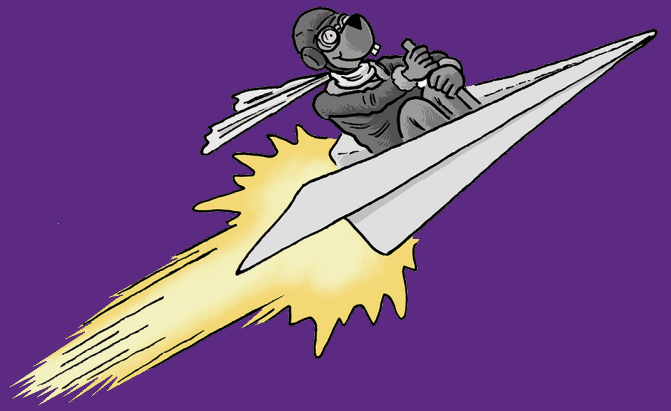
- extract all installed files from build system dependencies
- associate them with installation packages (so we know which SBOM to add this to)
- annotate & amend external dependencies (manual work, ideally requires upstreaming, and/or perhaps we're getting them over time)
- extract dynamic linking dependencies
  - known in the build system for internal components
  - external components might require `readelf` / platform-specific tools
- Extract static linking dependencies
  - known in the build system for internal components
  - manual work for externals – or can be provided with their SBOMs

Current state

- Collabora Online & core: static list of licenses available
- still a lot of work, but WIP patch here:



Conclusions & Questions



It sucks being at the top of the stack...
...and having a massive technology mix.

Get involved: We Are Hiring !

- https://collaboraonline.github.io/
- https://www.libreoffice.org/community/get-involved/

Hackfest

2 & 3 February 2026

Brussels

SIGN UP TODAY

Collabora Office