Copyright Statement Detection and Extraction Policy

Identification

We differentiate between authors, holders and copyright statements (as currently suggested by the ScanCode Toolkit).

- Authors are individuals or entities having written or contributed materials.
- Holders are individuals or entities holding rights on the materials.
- Copyright statements are markings of the authors/holders to identify holders and periods in which the materials have been created or published.

CIR-01: Copyright Identification Rule 01

As copyright statement we identify any statement that explicitly conveys the right-holder of the materials. A copyright statement is identified when marks such as

- ©,
- (C),
- (c),
- the term *copyright*, or
- variants of the above

are paired with

- names of individuals/entities, and/or
- time data (i.e. such as a year or a range of years).

CIR-02: Copyright Identification Rule 02

A plain mention in text (without a pairing as described in CIR-01) of a right-holder is not regarded an explicit copyright statement. We do not infer that the provider of the text intended to make an explicit statement here. Yet, the individual/entity is regarded a holder.

In case the plain mention in text is at a dedicated position, where a copyright statement is expected to be provided, the plain text mention is extracted as copyright statement.

AIR-01: Author Identification Rule 01

An author is identified as any individual or entity explicitly marked as having written or contributed to the materials. Markings for authorship include terms such as “written by,” “modified by,” “edited by,” or role identifiers like “original author”, “contributor” or “editor.” These attributions must refer clearly to participation in its creation. Individuals marked as “maintainers” or similar roles are not considered authors, as maintaining content does not necessarily imply authorship or original contribution.

HIR-01: Holder Identification Rule 01

A holder is identified when an individual or entity is referenced as owning, controlling, or holding rights over the material. Typical indicators include names in copyright statements or phrases such as “rights held by,” and “property of,”. The context of the statement must suggest ownership or legal authority over the material.

Copyright Extraction

CEP-01: Copyright Statements are extracted “as is”; they are not modified.

The copyright holder has made the statement. An extraction “as is” respects the choices and intentions of the copyright holder.

A technical limitation is different character sets and graphical^ representations. The extracted copyright statements must try to represent the original statement as far as possible.

Tabs and newlines are preserved.

Commenting related formatting (systematically trailing // or *) must not be preserved. These are regarded external boundary conditions.

CEP-02: “All rights reserved.” remarks are considered part of the copyright statement.

The “All rights reserved.” Statements are not considered part of the license. As such it is regarded part of the statement by the copyright owner. The statement is not required in general (since copyright-law is in place in participating countries). Yet. in case the copyright holder provides the remark, the remark is preserved as part of the copyright statement.

In general, all remarks by the copyright holder are regarded part of the copyright statement and are to be preserved.

CEP-03: A block of copyright statements is not decomposed.

A block of copyright statements may list several copyright holders of separate parts or changes. A copyright statement block may indicate a joint work of several copyright holders. A copyright statement may include “All rights reserved” remarks, not well aligned with the individual copyright holders.

The block is preserved since a decomposition requires additional knowledge on the copyrighted material. Preserving the block does not impose additional knowledge or interpret the intentions of the copyright holders. It does not disconnect individual marks from other parts.

CEP-04: Copyright statement on license texts must not be identified for the copyrighted materials or must be identified separately.

Licenses may have a copyright different from the materials supplied under the license. The copyright must be clearly assignable to the subject.

Copyrights of a license must not be ignored but identified as copyrights for the license.

CEP-05: Copyright statements are not consolidated.

I.e. a harmonization of copyright holders or consolidation/merges of time ranges are not allowed.

CEP-06: Copyrights are individually identified.

Copyrights are mapped to the associated license in a later processing step. A copyright may be mentioned several times with different license associations.

In the subsequent documentation it is sufficient to list a copyright statement once in combination with its license association. Repeating the same copyright with the same license is not regarded required.

CEP-07: URLs are considered part of the copyright statement.

When a URL appears directly adjacent to or within a copyright statement, it is considered part of that statement. This includes cases where the URL provides further information, points to the source, or represents a digital identifier. The URL is preserved in full to maintain the integrity of the statement as issued by the copyright holder.

Author Extraction

AEP-01: Authors are identified individually.

Authors are extracted as individual entities. If several authors are mentioned together in a list or block, each author is to be identified separately. Groupings or shared mentions must not result in a collective author identification. This ensures accurate attribution and supports disambiguation in later processing steps. Where multiple names are found, they are parsed and recorded as distinct individuals.

AEP-02: Contact information does not imply authorship.

The presence of contact information such as names, phone numbers or email addresses does not by itself indicate authorship or ownership. These references are often included for communication purposes only. Unless explicitly connected to an authorship or ownership claim, such information is not extracted as part of the author or holder data.

Holder Extraction

HEP-01: Holders are identified individually.

Similar to authors in CEP-07, holders are identified on a per-individual or per-entity basis. Even when several holders are named together in a single statement or block, each is treated as a separate holder. No merging or grouping is performed. This facilitates precise attribution and avoids misrepresentation of rights ownership.

General Extraction (apply to multiple targets)

GEP-01: Email addresses are considered part of the authors/holders.

Email addresses provided in proximity to authors or holders are preserved and associated with the respective entity. They are treated as identifying attributes and support disambiguation, especially in the presence of common names or organizational affiliations. Email addresses are normalized to lowercase and clearly marked by surrounding angle brackets (e.g. <user@example.com>). Obfuscated email addresses (e.g. user [at] example [dot] com) are not altered. Otherwise, email addresses are preserved in the context of the entity they relate to.

GEP-02: URLs associated to authors/holders are preserved.

Any URL associated with a holder or author (e.g., a GitHub profile link) may be used to help identify the individual or entity. For holders, the URL must appear

immediately following their name within the copyright statement. For authors, the URL must be clearly attributed to the individual.