# Software Heritage and IPOL,
# a fruitful collaboration towards reproducible research

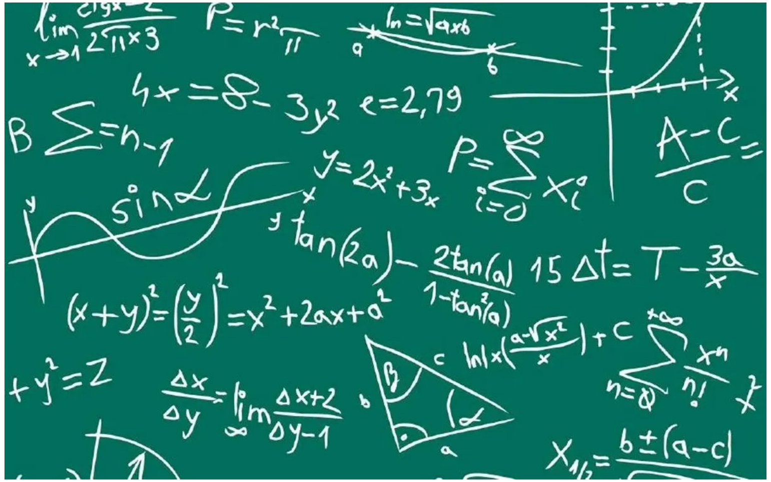
**Miguel Colom**
miguel.colom-barco at ens-paris-saclay d0t fr

## The origin of the IPOL journal

- **Started** in October **2009**
- The initiative of the **Image Processing Group** at CMLA (now **Centre Borelli**) at ENS-Cachan (now **ENS Paris-Saclay**)
- **First article** published in **2010**

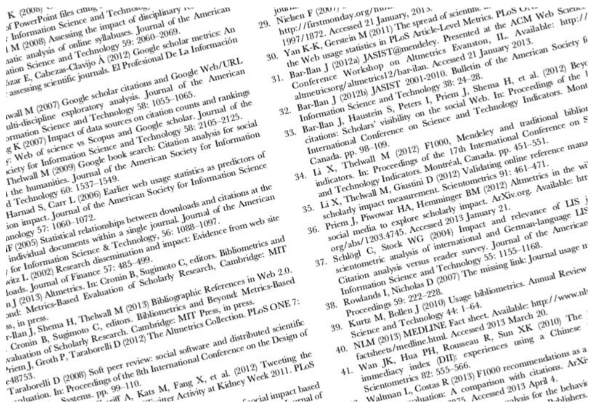
## IPOL's motivation

- The **reproducibility crisis** pointed out by Donoho et al
- Wanted to reveal the ***real*** **state of the art** in **image processing**
- Deeply understanding the methods. **All the mathematical details**
- **Reproducible research** :)



# How was it designed

- **Publication** = **article** + **source code** + **data** as a whole
  - no longer a PDF + supplementary material (code, data)
- Let **everybody test** the **algorithms**
  - Scientific community, students, industry, …
  - With the **own data**
- Make it **straightforward to use it**:
  - **Online demos**



K (2008)
of PowerPoint files citing Technology,
: Information Science and Technology
M (2008) Assessing the impact of disciplinary
matic analysis of online syllabuses. Journal of the American
ation Science and Technology 59: 2060–2069.
ózar E, Cabezas-Clavijo Á (2012) Google scholar metrics: An
: assessing scientific journals. El Profesional De La Información

wall M (2007) Google scholar citations and Google Web/URL
ulti-discipline exploratory analysis. Journal of the American
ormation Science and Technology 58: 1055–1065.
g K (2007) Impact of data sources on citation counts and rankings
y: Web of science vs Scopus and Google scholar. Journal of the
ciety for Information Science and Technology 58: 2105–2125.
Thelwall M (2009) Google book search: Citation analysis for social
the humanities. Journal of the American Society for Information
d Technology 60: 1537–1549.
Harnad S, Carr L (2006) Earlier web usage statistics as predictors of
ion impact. Journal of the American Society for Information Science
nology 57: 1060–1072.
F (2005) Statistical relationships between downloads and citations at the
individual documents within a single journal. Journal of the American
for Information Science & Technology, 56: 1088–1097.
witz L (2002) Research dissemination and impact: Evidence from web site
loads. Journal of Finance 57: 485–499.
n J (2013) Altmetrics. In: Cronin B, Sugimoto C, editors. Bibliometrics and
ond: Metrics-Based Evaluation of Scholarly Research, Cambridge: MIT
ss, in press.
r-Ilan J, Shema H, Thelwall M (2013) Bibliographic References in Web 2.0.
Cronin B, Sugimoto C, editors. Bibliometrics and Beyond: Metrics-Based
valuation of Scholarly Research. Cambridge: MIT Press, in press.
Priem J, Groth P, Taraborelli D (2012) The Altmetrics Collection. PLoS ONE 7:
e48753.
Taraborelli D (2008) Soft peer review: social software and distributed scientific
aluation. In: Proceedings of the 8th International Conference on the Design of
Systems. pp. 99–110.
iff A, Kats M, Fang X, et al. (2012) Tweeting the
itter Activity at Kidney Week 2011. PLoS
ocial impact based
urnal of

journ
29. Nielsen F (2007)
http://firstmonday.org/ht
1997/1872. Accessed 21 January, 2013.
30. Yan K-K, Gerstein M (2011) The spread of scientific
the Web usage statistics in PLoS Article-Level Metrics. PLoS ON
31. Bar-Ilan J (2012a) JASIST@mendeley. Presented at the ACM Web Scienc
Conference Workshop on Altmetrics Evanston, IL. Available: http://
altmetricsorg/altmetrics12/bar-ilan. Accessed 21 January 2013.
32. Bar-Ilan J (2012b) JASIST 2001-2010. Bulletin of the American Society f
Information Science and Technology 38: 24–28.
33. Bar-Ilan J, Haustein S, Peters I, Priem J, Shema H, et al. (2012) Beyo
citations: Scholars' visibility on the social Web. In: Proceedings of the
International Conference on Science and Technology Indicators. Mont
Canada. pp. 98–109.
34. Li X, Thelwall M (2012) F1000, Mendeley and traditional biblio
indicators. In: Proceedings of the 17th International Conference on S
and Technology Indicators. Montréal, Canada. pp. 451–551.
35. Li X, Thelwall M, Giustini D (2012) Validating online reference mana
scholarly impact measurement. Scientometrics 91: 461–471.
36. Priem J, Piwowar HA, Hemminger BM (2012) Altmetrics in the wi
social media to explore scholarly impact. ArXiv.org. Available: htt
org/abs/1203.4745. Accessed 2013 January 21.
37. Schlögl C, Stock WG (2004) Impact and relevance of LIS j
scientometric analysis of international and German-language LIS
Citation analysis versus reader survey. Journal of the American
Information Science and Technology 55: 1155–1168.
38. Rowlands I, Nicholas D (2007) The missing link: Journal usage m
Proceedings 59: 222–228.
39. Kurtz M, Bollen J (2010) Usage bibliometrics. Annual Review
Science and Technology 44: 1–64.
40. NLM (2013) MEDLINE Fact sheet. Available: http://www.nl
factsheets/medline.html. Accessed 2013 March 20.
41. Wan JK, Hua PH, Rousseau R, Sun XK (2010) The
immediacy index (DII): experiences using a Chinese
Scientometrics 82: 555–566.
Waltman L, Costas R (2013) F1000 recommendations as a
aluation: A comparison with citations. ArXi
875. Accessed 2013 April 4.
alysis for the behavi
Publishers.

## Structure of a publication

- The **PDF** of the article
- The **peer-reviewed source code**, under a **FOSS license**
- Any associated **data**

- An **online demo** (supplementary material)
  - An **archive** of **experiments**

## Let's take a look

https://ipolcore.ipol.im/demo/clientApp/demo.html?id=201&key=ED5EFD91CC1FF7A8A9B4F7305C901F48

## Particularities in the editorial process (1 / 2)

- **Not easy** to find **reviewers** both **experts** in the **scientific field** and the **implementation details** (code)
    - **Our solution**: always **consider two reviewers**
    - **One** more **focused** on the **scientific aspects** of the article and the **other** on the **code**

    - → **Can we really separate those two aspects?**
        - **NO**. They need to **work together**. We can't simply split the tasks. Both the **article** and the **code** are part of the **same publication**.

## Particularities in the editorial process (2 / 2)

- The **editors work** with the **authors** to improve their **code** until it's published
- We need **permanent identifiers** and **pointers** to the code **during the review process**
- **Also after publication**:
    - The **sources** need to be **preserved**. **Permanent** storage
    - The **identifiers** needs to follow a **standard**. FAIR data
    - One should be able to **cite** the **whole** or **pieces** of the source **code**
    - The **sources** need to be **referenced**, with **different granularity levels**

## General difficulties related to Reproducible Research

- The **source code** in the **author's website** could **disappear**
  - For example, a researcher moves to another university
- The **project in Github** could be made **private**
- **Github** could **close**! (See the precedent of **Google Code**)
- The **author** could **alter the history** and the **commit's tree**
  - Several **tools available**: BFG Repo-Cleaner, git filter-repo, …
- What about a **DOI?** Same problem: the pointed object **can be altered**. **Integrity not ensured**. **Responsibility** on **publisher's side**
- Each forge might provide their **own non-standard formats** for **referencing** the code
  - Probably **not the adequate granularity**
- **Not** an **standard** way to **cite software**
  - HAL, IPOL, and others the **include** the **SWHID** though

# Quick note: *intrinsic identifiers*

- Some **identifiers** are "**extrinsic**":
  - **Not computed** from the **object itself**
  - For example: **the DOI**
- **Intrinsic identifiers** are also **based on the contents of the object**
  - For example: the **SHA-1** sum of a file

# So… what do we need?

- A **repository** of **all source code**, with **perpetual archiving**
- A **dynamic** archive
  - If a new commits arrive, we want then in the stored copy
- **Traceability** and **complete metadata**
- **Identifiers** at different **granularity** levels
  - **Intrinsic**
  - Be able to **cite the sources** in a **standard** way
    - In France: good solution with HAL + Software Heritage for citation of code
- **Findable**: **from the identifier** we should **arrive to the archive itself**
- **Accessible** data: no registration, paywalls, …
- **Interoperable**: an **open specification** of the identifiers
- **Reusable**: identifiers and formats we can **apply** in **other contexts**

## So… any *good* solution around?

- Yes!
- **Software Heritage** provides **all we need** to **evaluate** and **publish reproducible research** and conduct **open science**
  - **Permanent storage**
  - **Intrinsic identifiers** (**SWHID**)
    - Granularity: snapshot, release, revision, directory, file, line, …
  - **Open standard**: SWHID standardization in progress…
  - Possibility to properly incorporate it within **software citations**
  - **No cost** for **authors** or **institutions** to **use** the platform

## Regular crawling

These software origins get continuously discovered and archived using the listers implemented by Software Heritage.

| | | |
|---|---|---|
| Bitbucket<br>2,539,527 origins | 56,983 origins | git<br>30,314 origins |
| R<br>26,984 origins | debian<br>136,866 origins | 54,628 origins |
| GitHub<br>205,730,285 origins | gitiles<br>10,232 origins | GitLab<br>4,245,668 origins |
| +++ git<br>3,267 origins | Gogs<br>197 origins | GO<br>1,076,337 origins |
| Guix<br>50,149 origins | GNU<br>354 origins | heptapod<br>1,232 origins |
| launchpad<br>512,270 origins | Maven<br>312,428 origins | NixOS<br>48,590 origins |
| npm<br>3,595,535 origins | 5,098 origins | Packagist The PHP Package Repository<br>305,886 origins |
| fedora PAGURE<br>67,596 origins | Phabricator<br>201 origins | pub.dev<br>50,994 origins |
| python Package Index<br>524,009 origins | SOURCEFORGE<br>381,373 origins | stagit<br>318 origins |

## Discontinued hosting

Discontinued hosting services. Those origins have been archived by Software Heritage.

| | | |
|---|---|---|
| GITORIOUS<br>122,014 origins | Google code<br>790,026 origins | Bitbucket<br>336,795 origins |

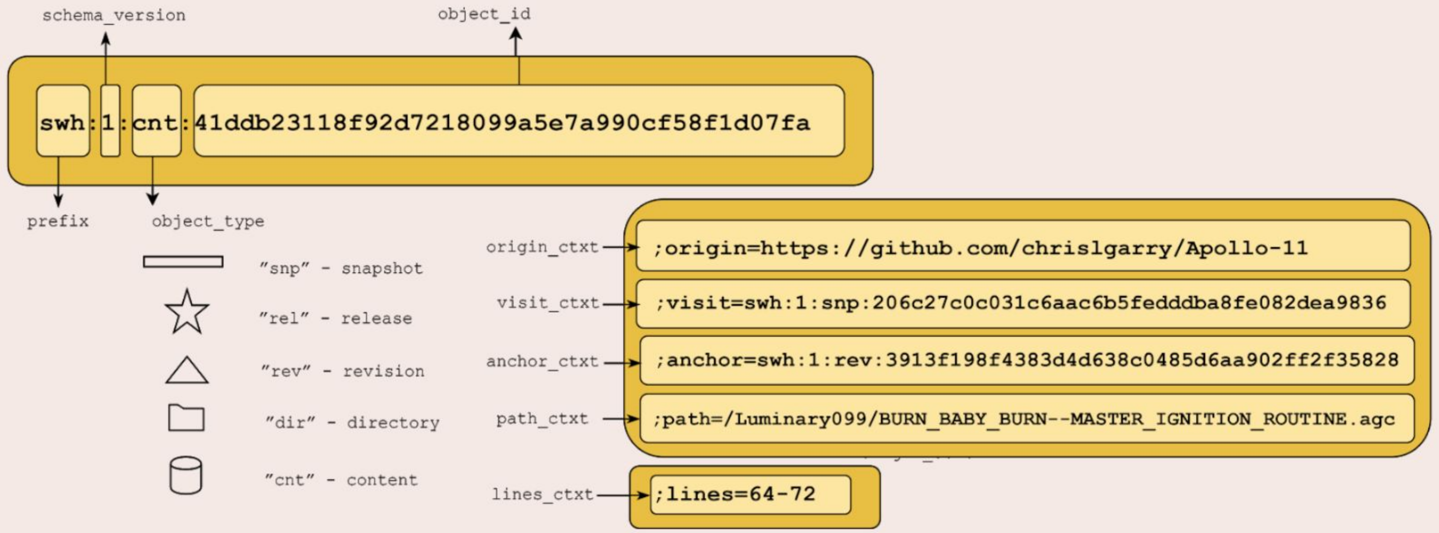
## On demand archival

These origins are directly pushed into the archive by trusted partners using the deposit service of Software Heritage.

| | | |
|---|---|---|
| eLife | HAL science ouverte | IPOL Journal |

## Structure of a SWHID identifier

link to full docs



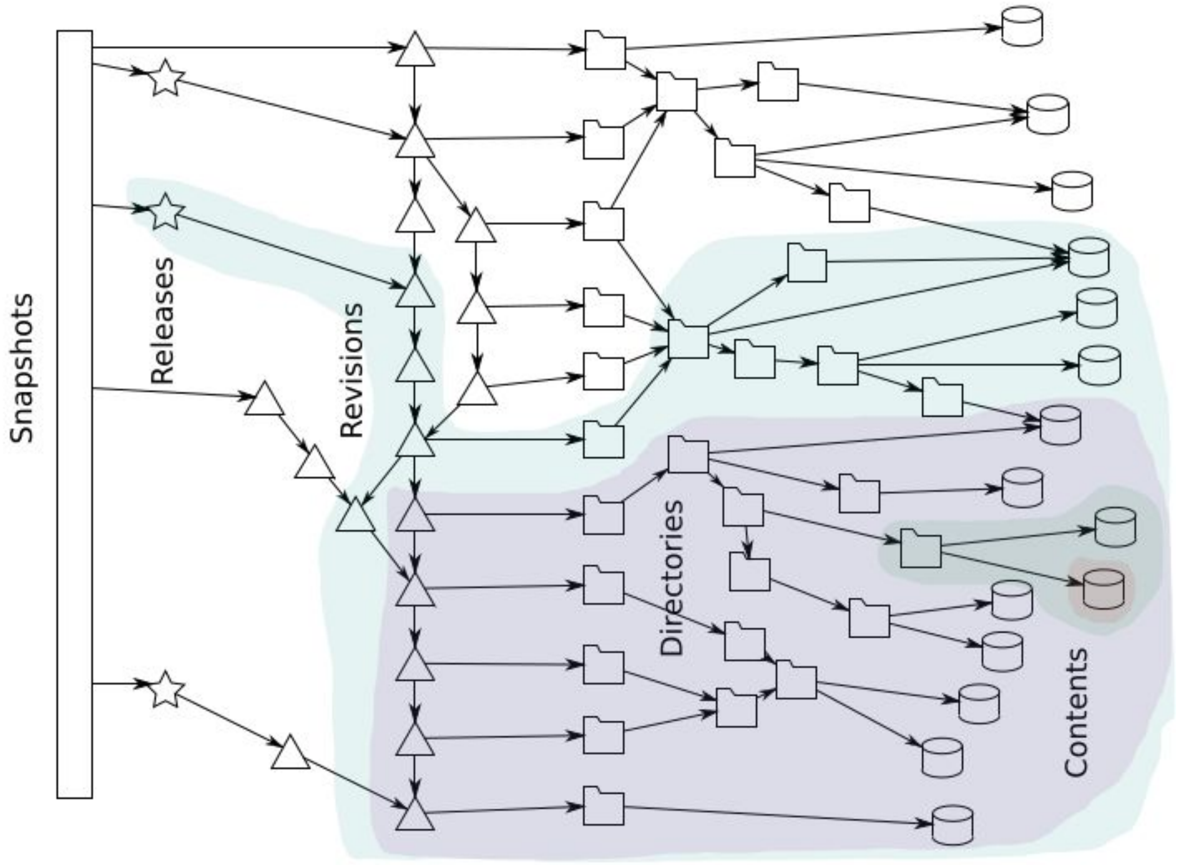
schema_version
object_id
swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa
prefix
object_type
"snp" - snapshot
"rel" - release
"rev" - revision
"dir" - directory
"cnt" - content
origin_ctxt
visit_ctxt
anchor_ctxt
path_ctxt
lines_ctxt
;origin=https://github.com/chrislgarry/Apollo-11
;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836
;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828
;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
;lines=64-72

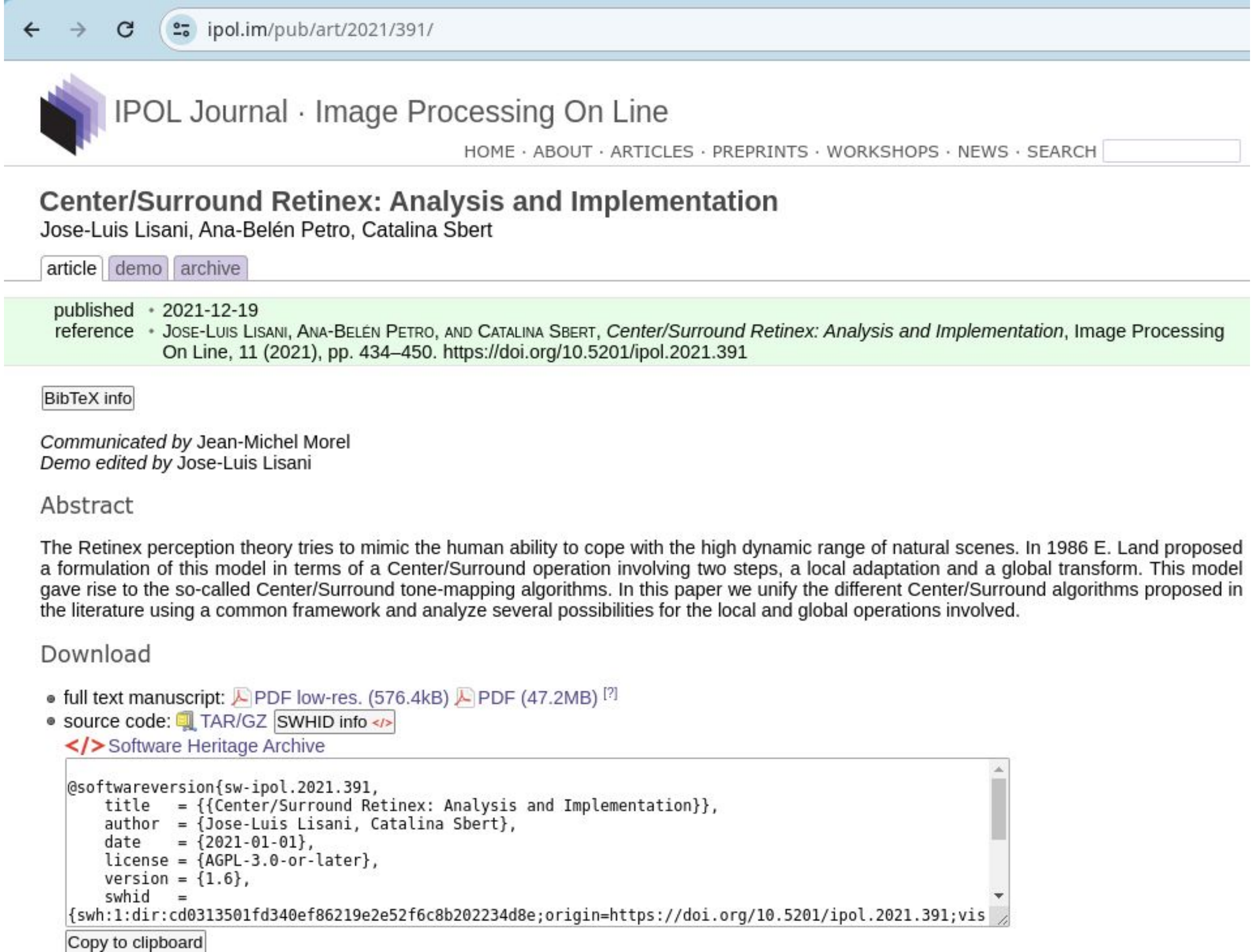
## Current status

30+B SWHIDs in the Software Heritage archive

Mention in Linux Foundation's SPDX 2.2; IANA registered; WikiData P6138

# How does it work? Merkle tree





ipol.im/pub/art/2021/391/

IPOL Journal · Image Processing On Line

HOME · ABOUT · ARTICLES · PREPRINTS · WORKSHOPS · NEWS · SEARCH

## Center/Surround Retinex: Analysis and Implementation

Jose-Luis Lisani, Ana-Belén Petro, Catalina Sbert

article demo archive

|  |  |
|---|---|
| published | 2021-12-19 |
| reference | Jose-Luis Lisani, Ana-Belén Petro, and Catalina Sbert, *Center/Surround Retinex: Analysis and Implementation*, Image Processing On Line, 11 (2021), pp. 434–450. https://doi.org/10.5201/ipol.2021.391 |

BibTeX info

*Communicated by* Jean-Michel Morel
*Demo edited by* Jose-Luis Lisani

### Abstract

The Retinex perception theory tries to mimic the human ability to cope with the high dynamic range of natural scenes. In 1986 E. Land proposed a formulation of this model in terms of a Center/Surround operation involving two steps, a local adaptation and a global transform. This model gave rise to the so-called Center/Surround tone-mapping algorithms. In this paper we unify the different Center/Surround algorithms proposed in the literature using a common framework and analyze several possibilities for the local and global operations involved.

### Download

- full text manuscript: PDF low-res. (576.4kB) PDF (47.2MB) [?]
- source code: TAR/GZ SWHID info </>
  </> Software Heritage Archive

```
@softwareversion{sw-ipol.2021.391,
    title   = {{Center/Surround Retinex: Analysis and Implementation}},
    author  = {Jose-Luis Lisani, Catalina Sbert},
    date    = {2021-01-01},
    license = {AGPL-3.0-or-later},
    version = {1.6},
    swhid   =
{swh:1:dir:cd0313501fd340ef86219e2e52f6c8b202234d8e;origin=https://doi.org/10.5201/ipol.2021.391;vis
```

Copy to clipboard

# IPOL's workflow

1. The **author develops** and versions **with git** with Gitlab, Github, or any other collaborative platform
2. When the **code is submitted**, the **editors** take note of the **submitted revision** (commit ID)
3. **IPOL** might **create** a **public git repository** for the code if not available (the authors might submit a ZIP file, for example)
4. The **authors** can **continue developing**, but **IPOL freezes** at that **particular** submitted **revision**
5. In case of **changes** (typically bug fixes), the editors can **merge** after **reviewing** and **update** the version under review
6. When the publication is **accepted**, it's **submitted to Software Heritage** for **archival**

## IPOL's code publication

- **When the code is accepted, it's submitted to Software Heritage for archival**
    - At this moment: **manual process** by the copyeditor

→ **We're working on improving (automating) this…**

# IPOL's code publication: ideas for the short term

- **Automatic deploy** to **Software Heritage**. Not only after publication, but also during the review process
- **Use** of **SWHIDs** in the review process, whenever they're available
- **Automatic download** of the **sources from a revision** of the **git repository**. **No** more (controlled) **packages** from a particular revision
- By default **prefer** the **copy** of the sources **in Software Heritage** instead of the local copy, whenever it's possible
- Allow for **integrity checks**. For example, given a file we could compute its hash, compare to an IPOL's database of published codes, find it, and obtain its SWHID along with all the metadata. **Traceability**.

- In short, we need to use and **develop new tools** to ensure: **findability**, **availability**, **integrity**, **traceability**, and better **reproducibility** .

# Conclusion

- The **objective of IPOL** is **communicating reproducible research** on **algorithms**, with **detailed mathematical descriptions** and providing the **source code** under a **FOSS license**. **Open science**.
- The **inclusion** of **Software** as part of the **publication** is **not trivial**
  - More **complex review process**
  - Needs to **reference properly** the **sources** during the review process and after publication.
  - Needs **permanent archival**
- **Software Heritage** has proved to be an **excellent ally** for **IPOL**, since it provides a **complete solution** and **infrastructure**
  - This was expected: **software is not supplementary material**, but a **main research artifact**. Reproducible research needs to that **Software** is properly **referenced**, **archived**, and **cited**. **Software Heritage fully covered** this need.

CENTRE BORELLI

Thank you for your attention!