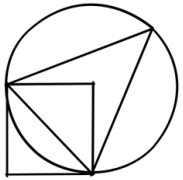
HAL et Grobid : structuration efficace à grande échelle des ressources à l'ère de l'IA

Luca Foppiano - luca@sciencialab.com



SciencíaLAB

Inria

Agenda

- Introduction
- Grobid in a few slides
- Grobid and LLM: comparison on fulltext extraction
- A glance at the roadmap
- The Inria Datalake project
- Conclusions
- How to stay in touch?

Introduction

- Worked with Patrice Lopez and Laurent Romary
- At inria from 2015 to 2019
- Started ScienciaLAB for working on R&D and maintenance of the Grobid family since 2024
  - Currently based in Portugal
  - Collaborating with Inria on several projects
- Focus on
  - Making the open source maintenance sustainable
  - Strengthening the community around Grobid
  - Take advantage of Large Language Models
- Looking for research collaborations

dfki ai

Helmholtz-Zentrum hereon

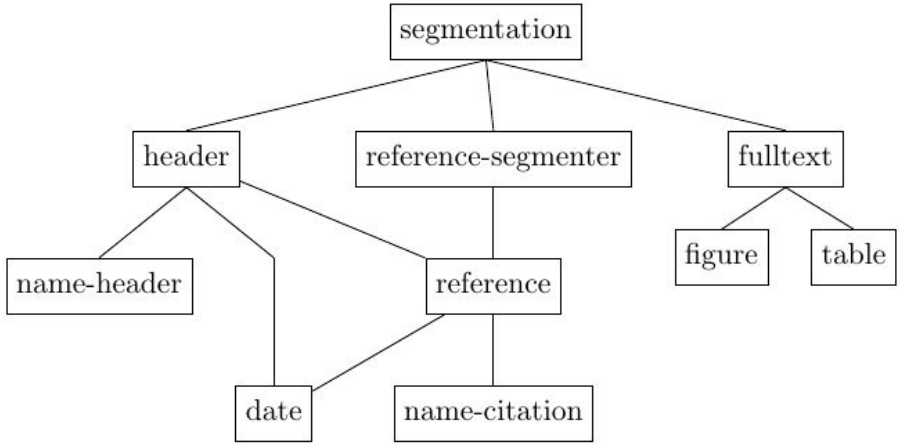
MAX-PLANCK-INSTITUT FÜR RECHTSGESCHICHTE UND RECHTSTHEORIE

Inria

Grobid in a few slides

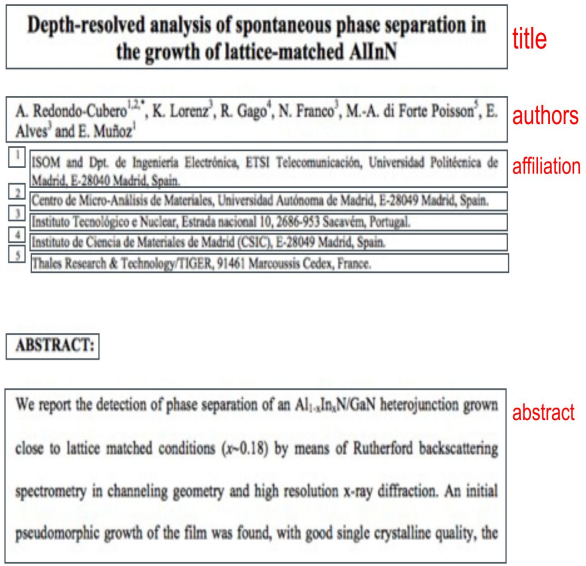
Grobid - Generation of Bibliographic data

- Grobid means GeneRation Of BIbliographic Data (started in 2008 by Patrice Lopez)
- Powered by several ML models applied in cascade
- Output XML TEI (Text Encode Initiative)
- Support standard layout of scientific articles
- Integrated in HAL for several years
- Last version: 0.8.2
- Next version: 0.9.0 (due Q1 2026)



Example: GROBID for meta-data extraction

GROBID (GeneRation Of Bibliographic Data) (*Lopez et al. 2015*)



Depth-resolved analysis of spontaneous phase separation in the growth of lattice-matched AlInN — title

A. Redondo-Cubero[1,2,*], K. Lorenz[3], R. Gago[4], N. Franco[3], M.-A. di Forte Poisson[5], E. Alves[3] and E. Muñoz[1] — authors

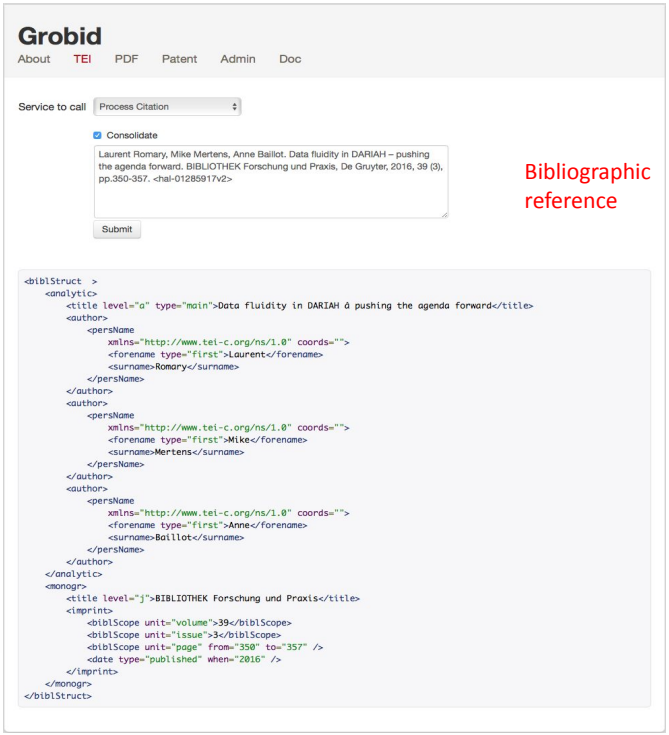
1 ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, E-28040 Madrid, Spain.
2 Centro de Micro-Análisis de Materiales, Universidad Autónoma de Madrid, E-28049 Madrid, Spain.
3 Instituto Tecnológico e Nuclear, Estrada nacional 10, 2686-953 Sacavém, Portugal.
4 Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.
5 Thales Research & Technology/TIGER, 91461 Marcoussis Cedex, France.

— affiliation

ABSTRACT:

We report the detection of phase separation of an $Al_{1-x}In_xN$/GaN heterojunction grown close to lattice matched conditions (x~0.18) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the

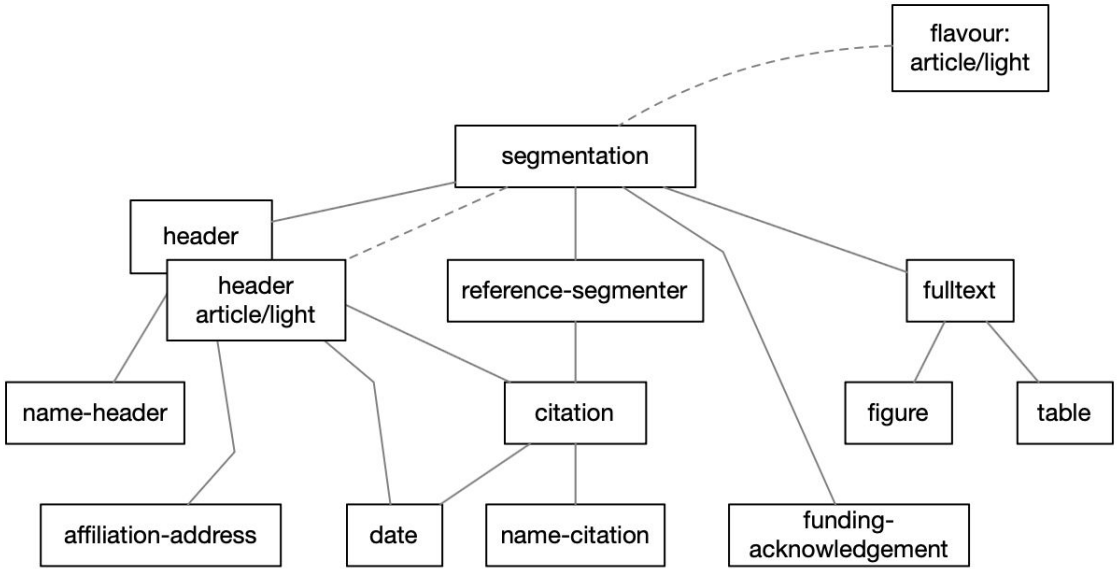
— abstract



Slide from Laurent Romary - Grobid Camp 2025

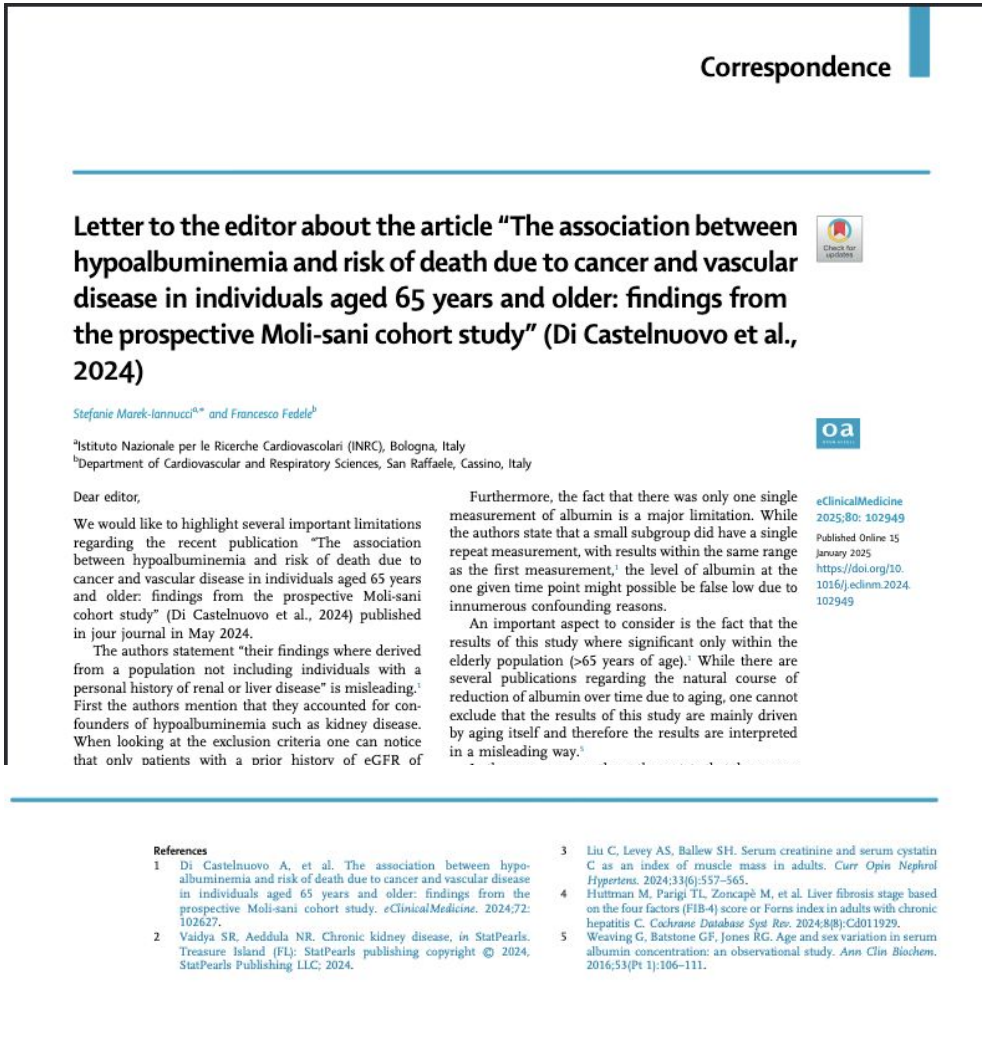
Grobid flavours

- Feature from Grobid 0.8.2
- Any model can be overridden with minimal changes
- Support new type of document (e.g. FDI regulation reports, special reports, etc..)
- Iterative development and evaluation of new document types



Grobid flavours examples

- **article/light**: simple header structure (title, authors, pub date), no tables, no figures
- **article/light-ref**: same as article/light with references
- **SDO/IETF**: standard documents (e.g. wifi, 5g, etc..)
- **Law and History** (WIP): support of documents with references as footnotes



Correspondence

Letter to the editor about the article "The association between hypoalbuminemia and risk of death due to cancer and vascular disease in individuals aged 65 years and older: findings from the prospective Moli-sani cohort study" (Di Castelnuovo et al., 2024)

Stefanie Marek-Iannucci[a,*] and Francesco Fedele[b]

[a]Istituto Nazionale per le Ricerche Cardiovascolari (INRC), Bologna, Italy
[b]Department of Cardiovascular and Respiratory Sciences, San Raffaele, Cassino, Italy

Dear editor,

We would like to highlight several important limitations regarding the recent publication "The association between hypoalbuminemia and risk of death due to cancer and vascular disease in individuals aged 65 years and older: findings from the prospective Moli-sani cohort study" (Di Castelnuovo et al., 2024) published in jour journal in May 2024.

The authors statement "their findings where derived from a population not including individuals with a personal history of renal or liver disease" is misleading.[1] First the authors mention that they accounted for confounders of hypoalbuminemia such as kidney disease. When looking at the exclusion criteria one can notice that only patients with a prior history of eGFR of

Furthermore, the fact that there was only one single measurement of albumin is a major limitation. While the authors state that a small subgroup did have a single repeat measurement, with results within the same range as the first measurement,[1] the level of albumin at the one given time point might possible be false low due to innumerous confounding reasons.

An important aspect to consider is the fact that the results of this study where significant only within the elderly population (>65 years of age).[1] While there are several publications regarding the natural course of reduction of albumin over time due to aging, one cannot exclude that the results of this study are mainly driven by aging itself and therefore the results are interpreted in a misleading way.[5]

eClinicalMedicine 2025;80: 102949
Published Online 15 January 2025
https://doi.org/10.1016/j.eclinm.2024.102949

References
1 Di Castelnuovo A, et al. The association between hypoalbuminemia and risk of death due to cancer and vascular disease in individuals aged 65 years and older: findings from the prospective Moli-sani cohort study. *eClinicalMedicine*. 2024;72:102627.
2 Vaidya SR, Aeddula NR. Chronic kidney disease, *in* StatPearls. Treasure Island (FL): StatPearls publishing copyright © 2024, StatPearls Publishing LLC; 2024.
3 Liu C, Levey AS, Ballew SH. Serum creatinine and serum cystatin C as an index of muscle mass in adults. *Curr Opin Nephrol Hypertens*. 2024;33(6):557–565.
4 Huttman M, Parigi TL, Zoncapè M, et al. Liver fibrosis stage based on the four factors (FIB-4) score or Forns index in adults with chronic hepatitis C. *Cochrane Database Syst Rev*. 2024;8(8):Cd011929.
5 Weaving G, Batstone GF, Jones RG. Age and sex variation in serum albumin concentration: an observational study. *Ann Clin Biochem*. 2016;53(Pt 1):106–111.

Grobid and LLMs

Document structuring with LLM

- The LLM (Large Language Models) are becoming cheaper and more effective
- A second category called vLLM (Visual Language Models) consists of language models with visual channel (they process images)
- However, (v)LLM popularity is given by perception rather than data
- For document structuring, standard LLM work well only specific tasks
- Grobid performs multiple structuring tasks at very low price, e.g.
  - Authors/ Affiliation extraction
  - Reference extraction and structuring
  - Fulltext extraction
  - …
- We tried to assess how vLLM perform against Grobid

Fulltext extraction comparison of Grobid and vLLM

- Using a serverless infrastructure (modal.com) for running LLMs (billed by second of GPU use)
- Evaluate vLLM with less than 2B parameters
- Comparison of costs, time and accuracy, on tiny dataset (10 documents)
- Need to manually craft new dataset
- Existing dataset (Olmobench, OmniBench, DotOCRBench) are page-based.
- **Results from a 2000 PMC documents benchmark**

Model	GPU	Documents	Avg Runtime per doc	Runtime per million docs	Total Cost (USD)	Cost per million docs (USD)
DotsOCR	A100-40GB	1943	12s (estimated)	138 days	15$	7500
OlmoOCR	A100-40GB	1943	13s (estimated)	150 days	14$	7450
Docling	A100-40GB	1943	7s	81 days	6$	**3000**
GROBID	CPUs	1943	**2.5s**	28 days		

Fulltext extraction comparison with vLLM

- **NED**: Normalized Edit Distance (Levenshtein distance)
- **ROUGE**: evaluate the quality of text generated by natural language processing models. Emphasizes recall by measuring the overlap of n-grams between the two texts.
- **Reading order (header level):** check the sections to be in the correct order (coarse)
- **Reading order (paragraph level):** check the paragraph to be in the correct order (granular)

Metric	GROBID	DotsOCR	Docling	Olmocr
Avg NED (Higher is better)	0.7627	0.7051	**0.7716**	0.7847
Avg F1 (Higher is better)	0.8945	0.8592	0.8773	**0.8962**
Avg ROUGE-L (Higher is better)	0.8983	0.8493	0.9035	**0.9131**
Avg Reading Order Header Level(Higher better)	**0.5192**	0.0858	0.2961	0.2909
Avg Reading Order paragraph Level (Higher better)	**0.68**	0.0858	0.668	0.5446
Avg Coverage (Higher is better)	0.7875	0.7396	0.7852	**0.8139**
Documents Evaluated	1943	1943	1943	1942

Limitations and Challenges

Grobid quality is on pair with vLLM, costs are preferred for large scale processes

Notes:

- It evaluates only the fulltext
- Most vLLM output the text as it’s appearing in the document
- vLLM would help solving PDF encoding issues (e.g. badly formatted formulas)
- Grobid focuses on providing a structured logical document
  - Pages, break line are meaningless
  - Header information stays in the header, even if the publisher put them at the end of the document
- We are working towards an alignment between “OCRed” output from vLLM and Grobid structure

Glance at the roadmap

Grobid Camp 2025

In November 2025, with the support of INRIA and the MESR we held the “Grobid Camp”:

- Two days of exchanges
- Gathering the main institutions in France (INRIA, MESR, ABES, ISTEX, MATILDA, and many more..)
- Work on a proposal Roadmap based on use cases
- Plan to organise another Grobid Camp in 2026

Elements of a Grobid roadmap

- **Landscaping usages – community management**
  - **Connecting all projects using Grobid**
  - **Communication**
  - **Documentation**
  - **Ateliers**
- Improving core capabilities
  - Multilingualism (CJK)
  - Robust first level segmentation model, visual models
  - Problematic PDFs (encodings, OCR)
- **Management of new models**
  - **Conflicts of interest, material and methods, tableaux, figures, licences**
  - **Books/Theses, grey literature, preprint (non standard)**
- Grobid architecture
  - Orchestration d'autres modèles
- Peripheric support activities
  - Data model, format conversion (JATS)
  - Light-weight results, json
  - Scores
- Various levels of contribution
  - Contributing to the source code
  - Defining new models
  - **Contributing test and training data (golden set)**
  - **Synthetic data**
  - **Contributing issue**
  - Sharing computing facilities
- **Services**
  - **Grobid-online, where?**
  - **Simple use case deployment**
  - **Sharing results, IP issues**
- Pooling financing
  - Eliciting Grobid in funding applications
  - Towards a Grobid foundation?
- Benchmarking
  - LLMs, etc.

Slide from Laurent Romary - Grobid Camp 2025

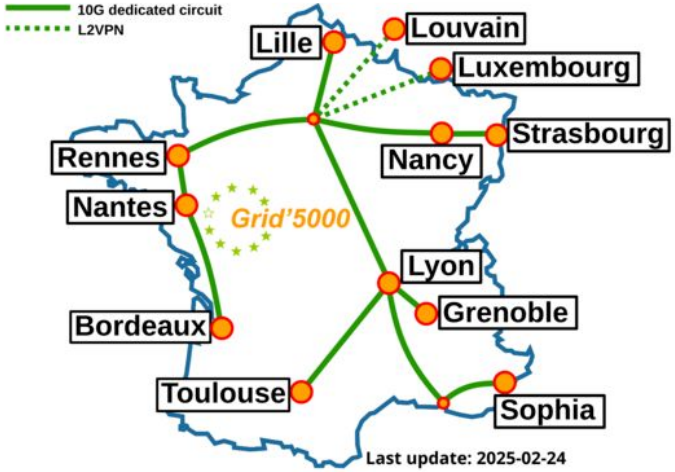
The Inria Datalake project

Objective

- Provide a shared large data pool with data from HAL at different processing stages
  - Structured documents (Grobid)
  - Specialized processing:
    - Software mentions
    - Dataset mentions
    - Etc..
- The main advantages are:
  - Reduce the effort to process the same data
  - Align improvements and data versions
  - Collect feedback in a single place

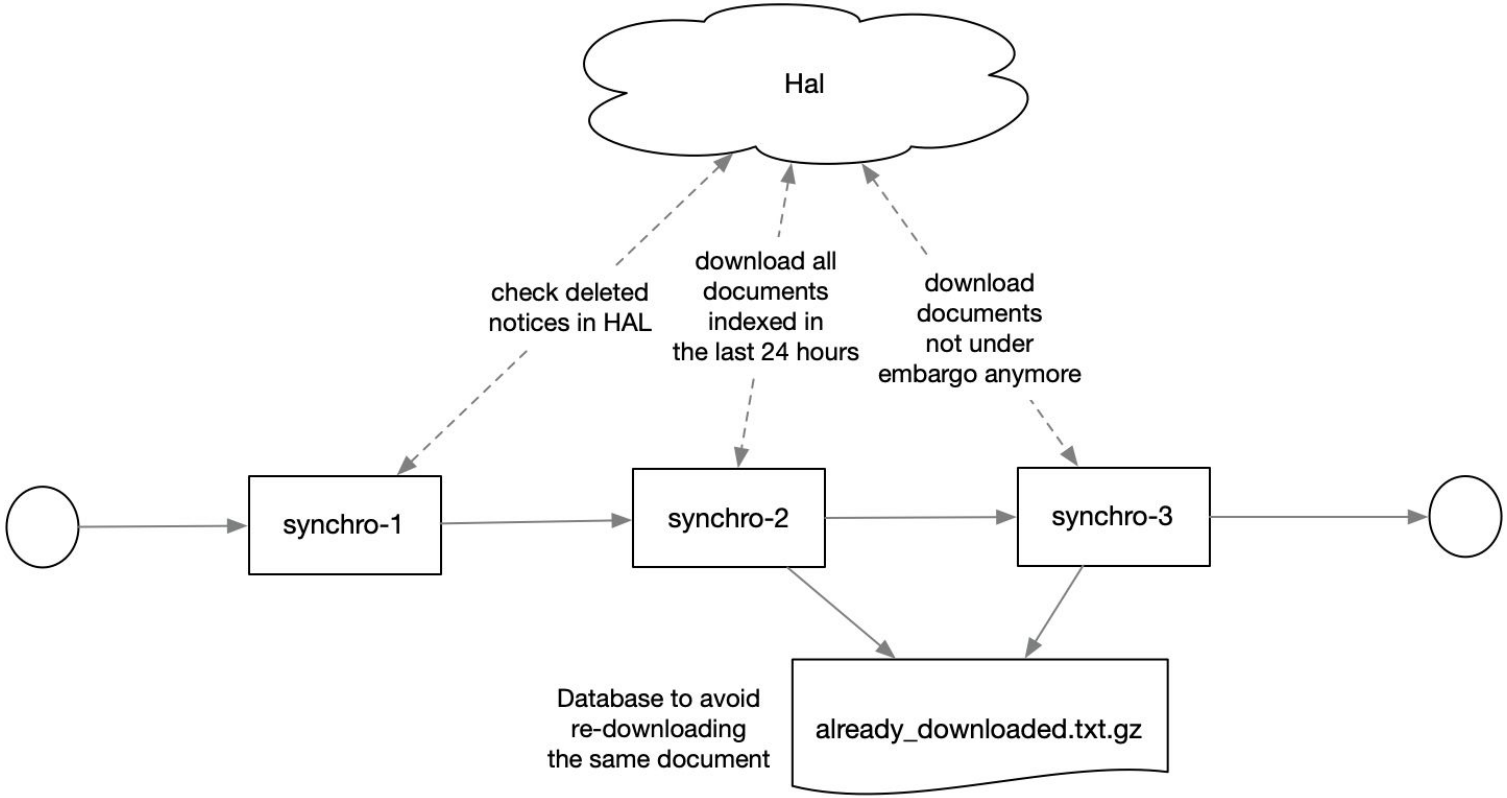
Processing

- HAL Synchronization
  - Runs on a shared distributed and large scale infrastructure: Grid5000
  - Runs every night
  - Fetch a list of all the new registrations in HAL
  - Download the metadata of the new documents
  - Download the PDF (when available)
  - Skip data protected by “embargo” (not yet publicly available)
- PDF Processing
  - Grobid process PDF to TEI-XML
  - PDF are removed after
- Specialized processes
  - Run asynchronously
  - Software mention extraction from TEI-XML

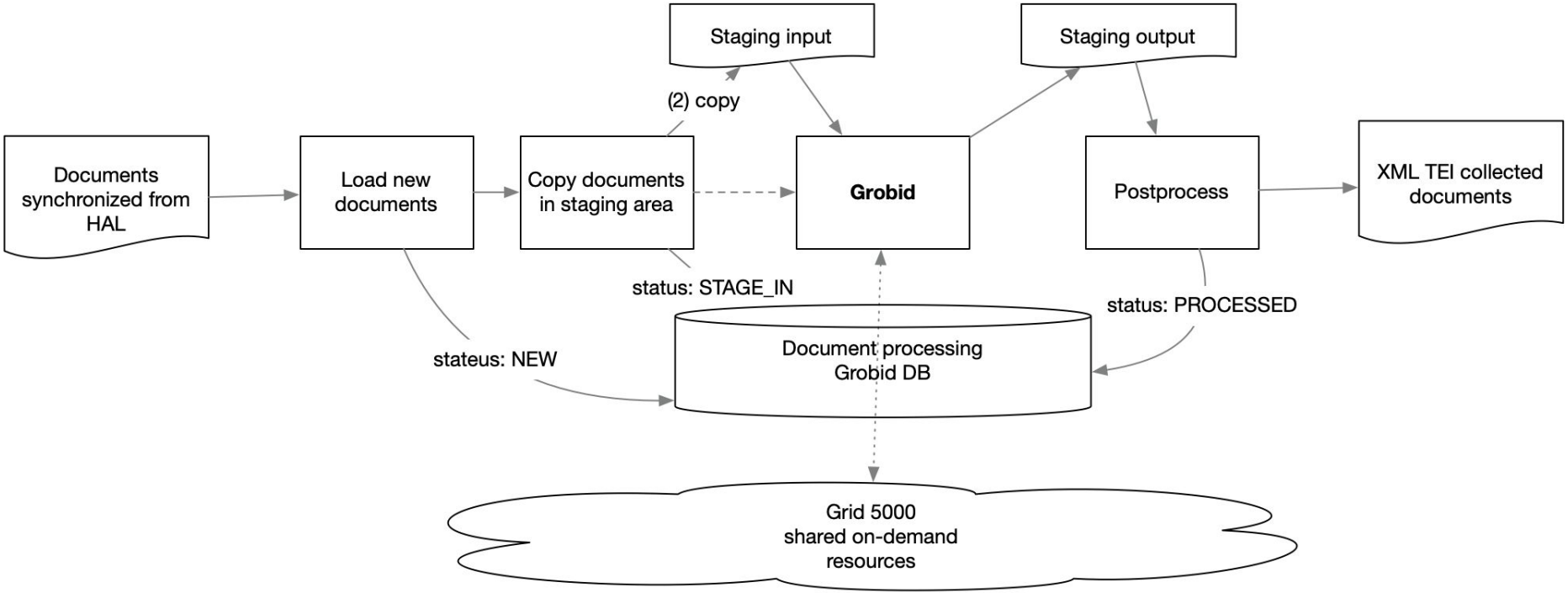


Grid’5000 is a large-scale, open research infrastructure designed for experimentation in distributed, parallel, and cloud computing.

HAL Synchronisation



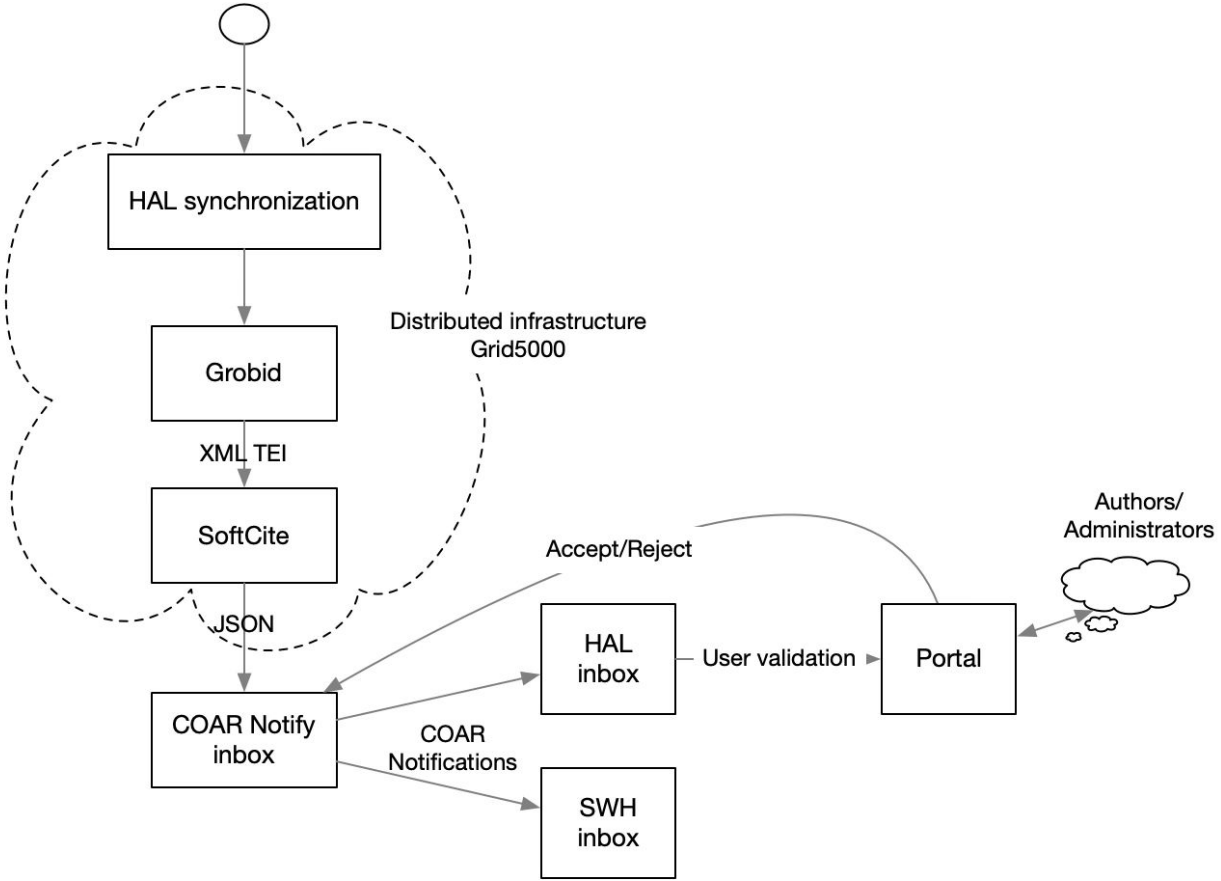
PDF processing



Use case: Software mentions validation

- Scope of the EU project: SoFAIR
- SoFAIR aims to improve extraction of software mentions from scientific articles
- Software mention extraction require validation
  - Homonyms
  - Name changing, rebranding
  - False positives
- Involving both HAL and Software Heritage
- Currently in preprod - to be rolled out in 2026

Use case: Software mentions validation



Status

- We tested the infrastructure, processed 1.6M documents from HAL, in one week
- Currently, the daily jobs are under tests (SoFAIR use case)
- The Grid 5000 / ABACA infrastructure provide a very cost-efficient infrastructure
- We plan to start releasing the processed data during 2026

Conclusions

- Maintenance on Grobid will continue
- Plan to strengthen the community
- Plan to align structure and data quality from vLLM
- Roadmap for the following years to come
- Inria Datalake project to share resources
  - Providing extracted document to research teams
  - To share resources for new specialised processes

How to stay in touch?

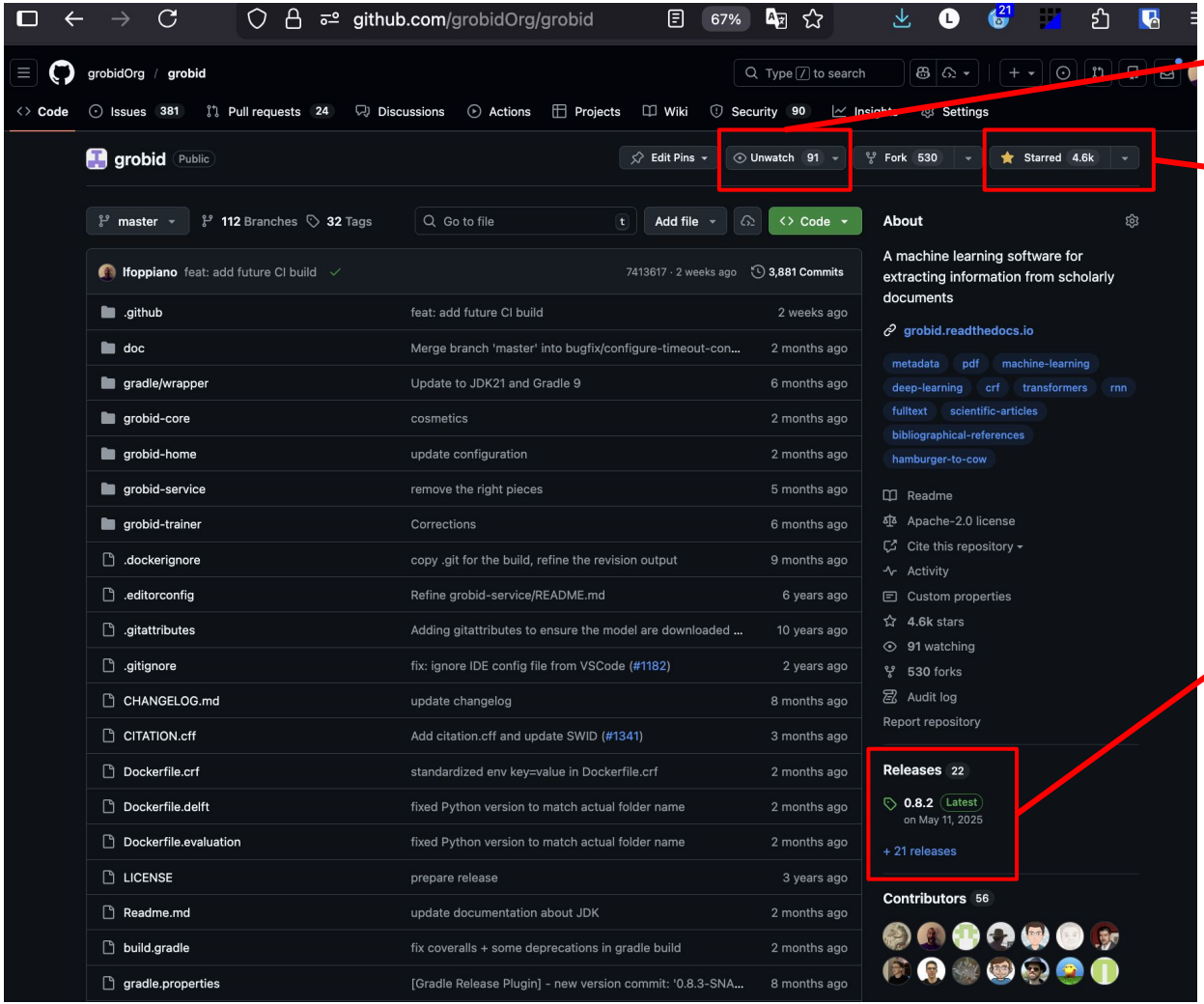
How to stay in touch

Grobid: https://github.com/grobidOrg/grobid (organisation https://github.com/grobidOrg)

Datalake: https://github.com/Inria-Datalake

Support:

- INRIA (French National Institute for Research in Digital Science and Technology)
- MESRE (French Ministry of Higher Education, Research and Innovation)
- And many others…



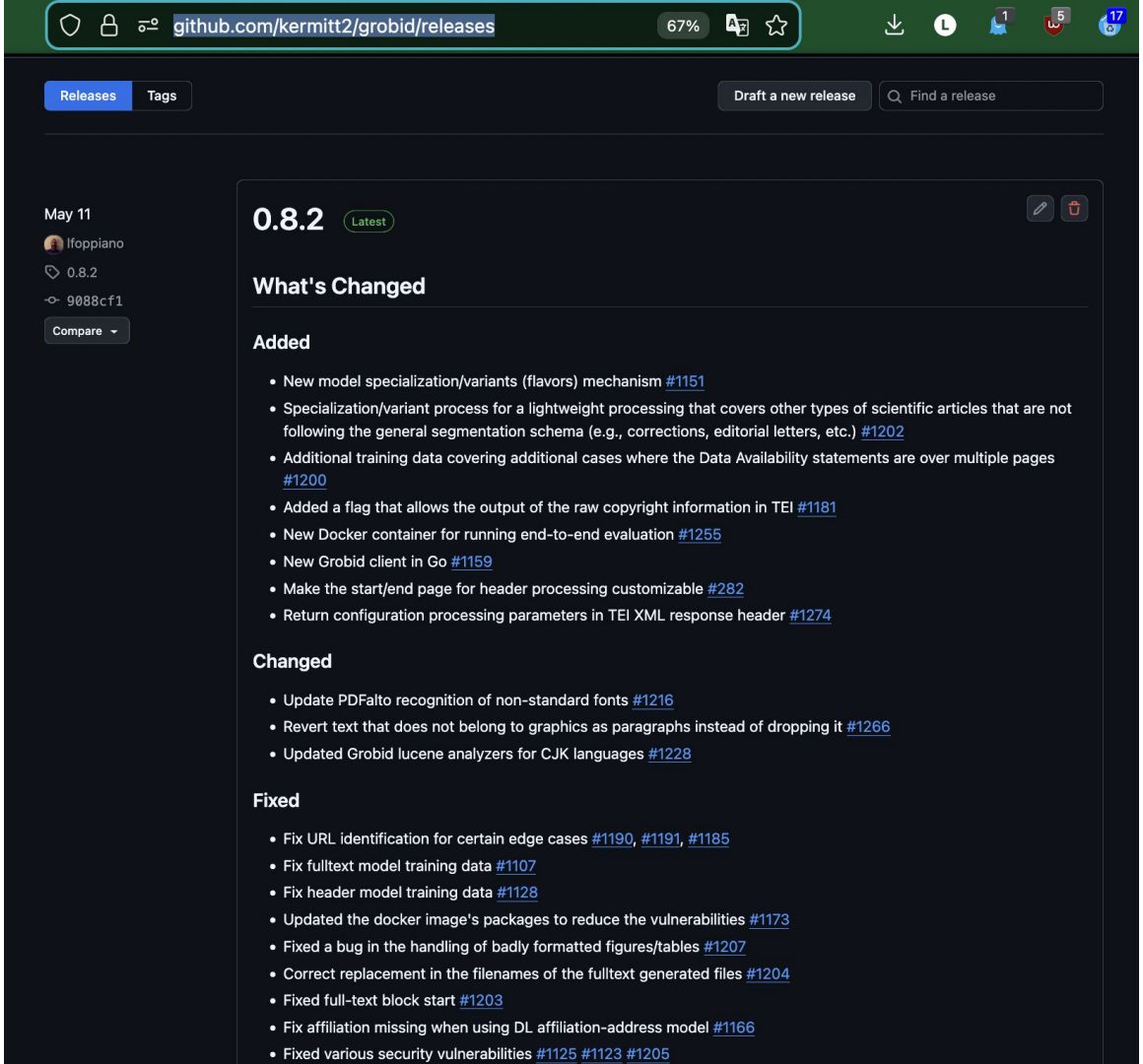
Get updates

Give support :-)

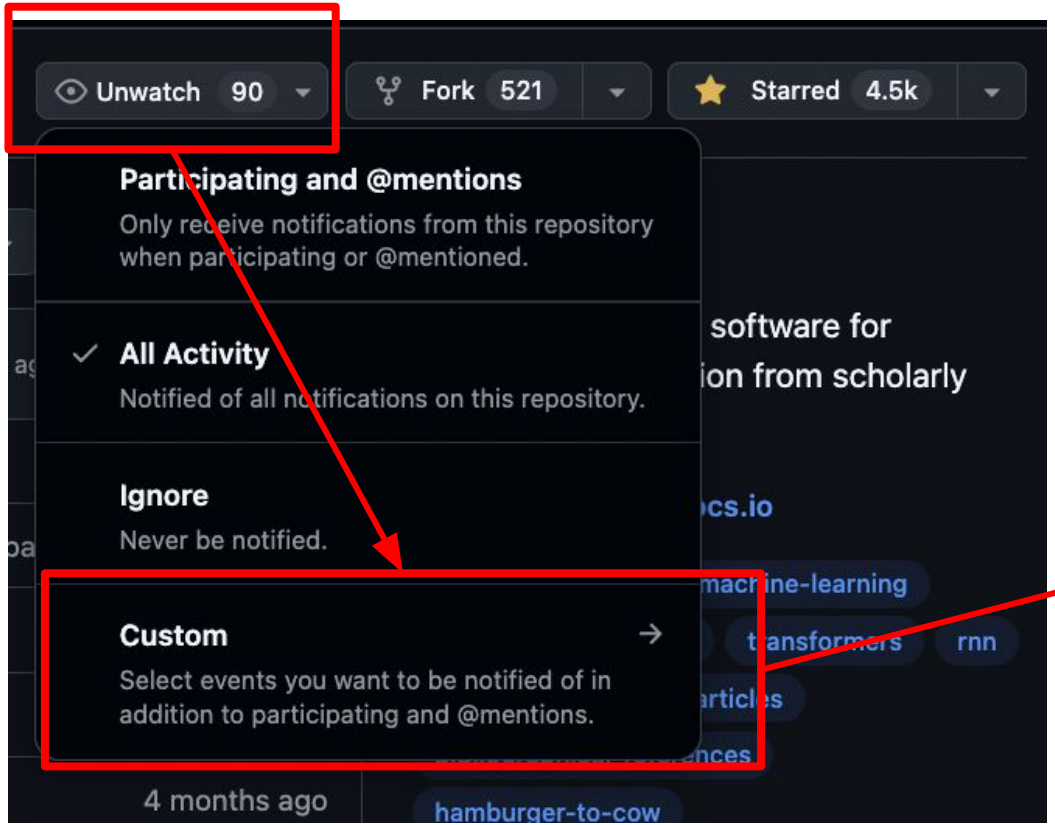
Release information

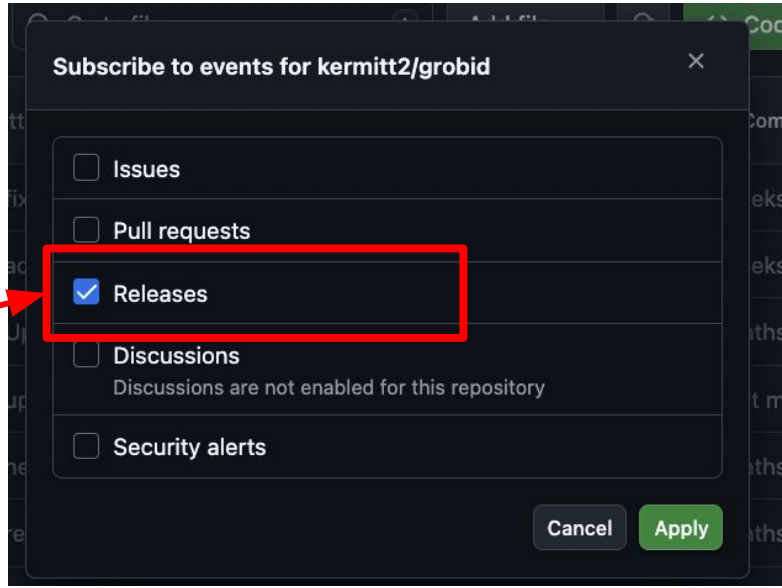
Release information

One release occurs approximately every 6-8 months.

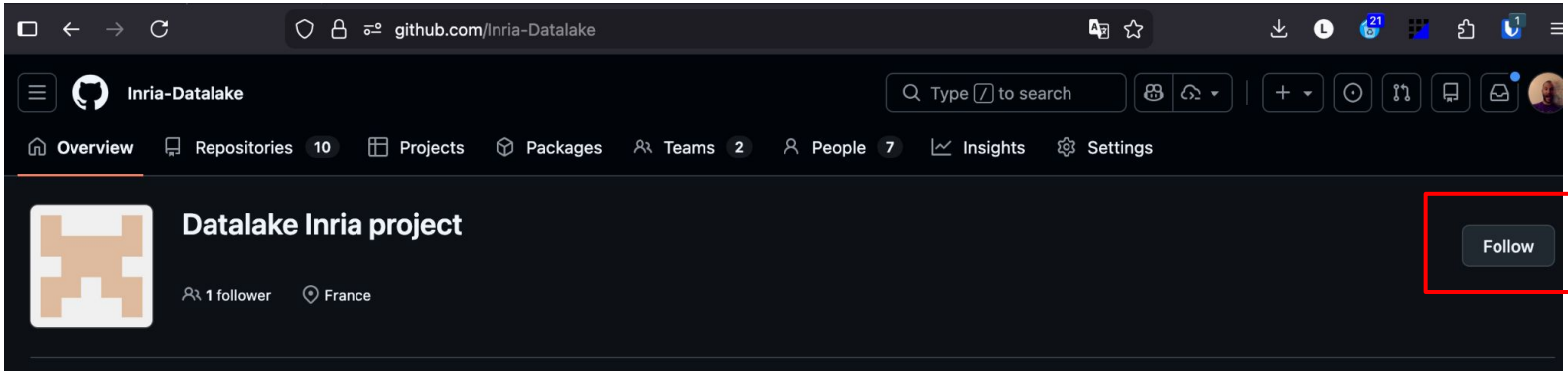


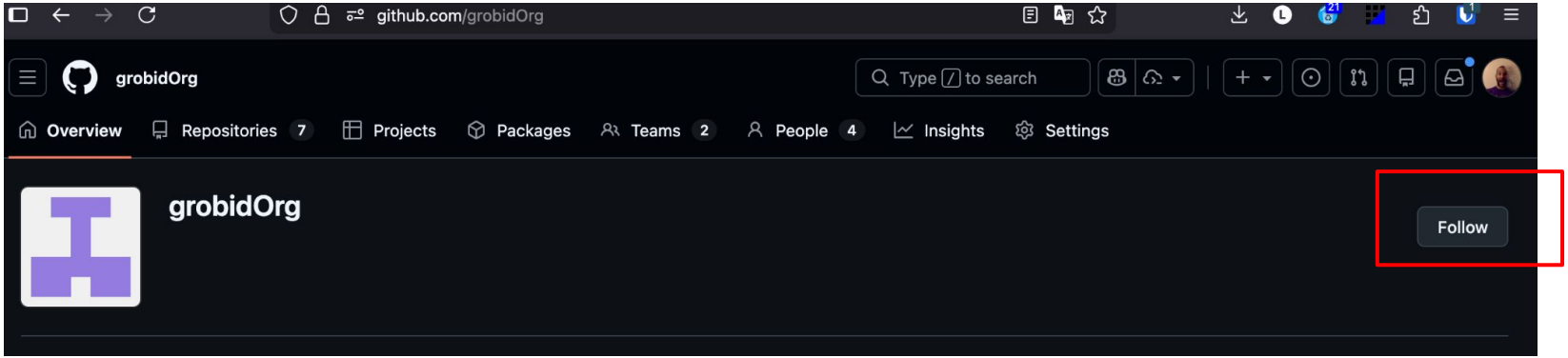
How to get notified of new releases?



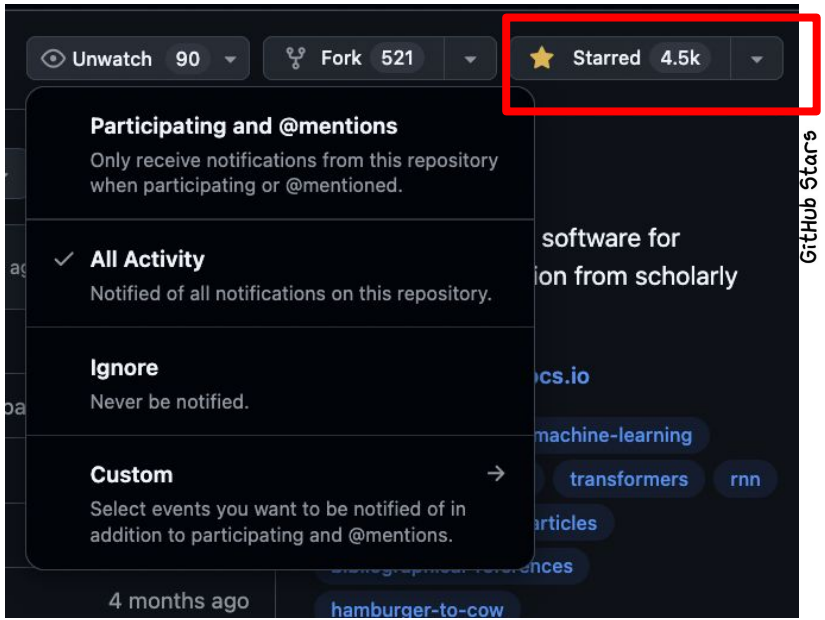


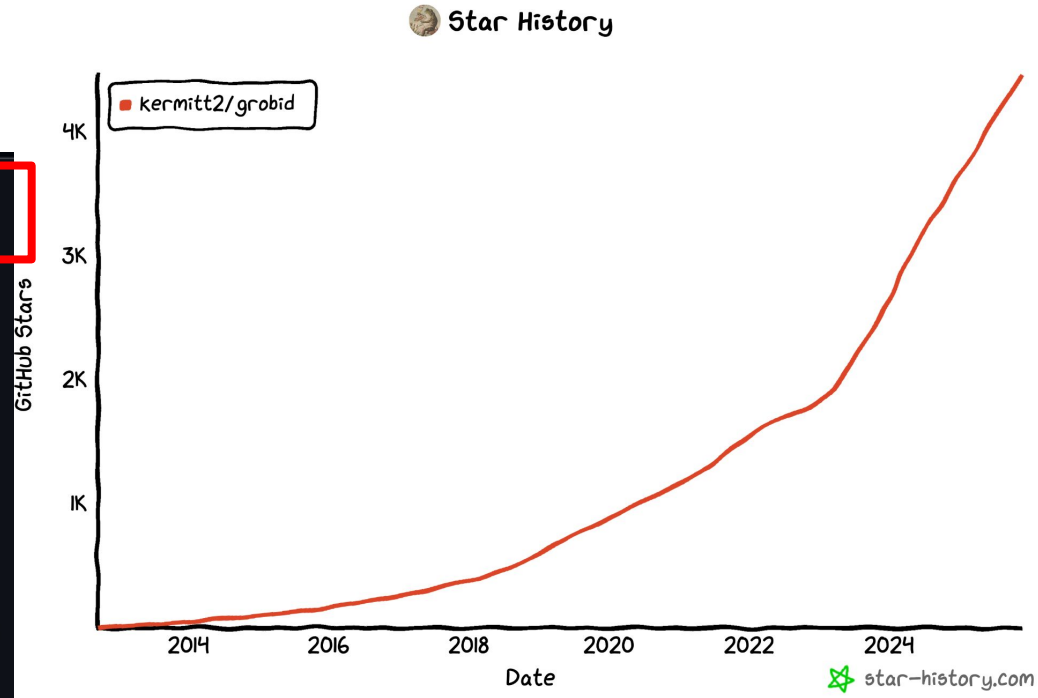
Follow the organisations





Optional, star the Grobid project! :-)





Thank you