



Foto: Freepik.com. Reprodução.

O Software Heritage visa construir o maior repositório universal de código-fonte aberto disponível – uma infraestrutura global essencial para a Ciência Aberta.

Software Heritage

por Roberto di Cosmo, Fernanda Antonia da Fonseca Sobral e Claudia Bauzer Medeiros

Resumo

Roberto di Cosmo, professor da Universidade de Paris Diderot e líder do Software Heritage, destaca-se pela defesa do software livre e pela preservação do conhecimento digital. Em entrevista, ele aborda a importância da Ciência Aberta, ressaltando como o Software Heritage pode fortalecer a pesquisa global e o papel do Brasil nesse cenário.

Palavras-chave: Software Livre, Ciência Aberta, Preservação Digital, Software Heritage

Roberto di Cosmo é uma figura de destaque no mundo da ciência da computação e da defesa do software livre. Professor titular de Informática na Universidade de Paris Diderot desde 1999, ele também preside o conselho de

administração e o conselho científico do Instituto IMDEA Software, além de liderar o capítulo de software do Comitê Nacional para a Ciência Aberta na França. Atualmente vinculado ao Inria, um dos maiores institutos de pesquisa em ciência

da computação do mundo, Roberto di Cosmo tem uma trajetória marcada por contribuições significativas em áreas como informática teórica, programação funcional, semântica de linguagens de programação e sistemas de tipos. Sua paixão

pelo software livre não é recente (Figura 1). Desde 1998, ele tem sido um defensor ativo desse movimento, contribuindo com obras influentes como o best-seller “Hijacking the World”, além de seminários, artigos e desenvolvimento de software. Em 2007, criou o grupo temático de software livre da Systematic, que financiou mais de 50 projetos de pesquisa e desenvolvimento em código aberto, com um investimento consolidado de mais de 200 milhões de euros. Entre 2010 e 2018, dirigiu o IRILL, um centro de pesquisa dedicado à qualidade do software livre, consolidando sua reputação como um líder na área. Um de seus projetos mais ambiciosos é o Software Heritage, iniciativa lançada em 2015 que ele ainda dirige. Em colaboração com a Unesco, o Software Heritage tem como objetivo construir o maior repositório universal de código-fonte disponível publicamente. Essa iniciativa não apenas preserva o conhecimento digital, mas também se torna uma infraestrutura global essencial para a Ciência Aberta, integrando-se a editoras, repositórios de pesquisa e iniciativas acadêmicas. Para o Brasil, a participação no Software Heritage representa uma oportunidade única de liderar em Ciência Aberta e preservação digital. Ao garantir que o software de pesquisa produzido no país seja preservado e acessível, o Brasil pode fortalecer sua posição no cenário científico global e assegurar que o conhecimento gerado hoje continue a beneficiar as gerações futuras. Nesta entrevista, realizada e traduzida por Claudia Bauzer Medeiros,

professora do Instituto de Computação da Universidade Estadual de Campinas (Unicamp), Roberto di Cosmo compartilha suas visões sobre software livre, Ciência Aberta e o papel transformador do Software Heritage na preservação do conhecimento digital.

Ciência & Cultura – O que é o Software Heritage e por que a Unesco o financia desde 2017, considerando-o como uma infraestrutura chave para preservar software como parte do patrimônio mundial da humanidade?

Roberto di Cosmo – O software é um pilar fundamental do conhecimento moderno. Ele impulsiona descobertas científicas, alimenta infraestrutura crítica e molda a evolução tecnológica. No entanto, ao contrário de publicações e dados científicos, o software continua sofrendo o risco de desaparecer por razões de obsolescência, falta de arquivamento adequado ou mudanças organizacionais. A Software Heritage foi criada para endereçar este desafio. É um sistema universal de arquivamento a longo prazo de código-fonte de software, garantindo que todo código disponível seja publicamente coletado, preservado e tornado acessível para as gerações futuras. Ao prover uma infraestrutura aberta, replicada e padronizada, a Software Heritage protege não apenas o código em si, mas também o conhecimento embutido nele. Desde que iniciou suas atividades, a Software Heritage coletou mais de 340 milhões de projetos de software, num total de 22 bilhões de arquivos únicos, sem

duplicação. Ao reconhecer a importância do software no ecossistema científico e cultural, a Unesco se tornou parceira da Software Heritage como parte de sua estratégia global para preservar o conhecimento digital. Tawfil Jelassi, Diretor Geral Assistente do Setor de Comunicação e Informação da Unesco, expressou muito bem este conceito no seu discurso durante o Simpósio Anual do Software Heritage em 2023:

“O código-fonte do software representa um conhecimento único da história recente da humanidade. É crucial trabalhar de forma coletiva, para que o conhecimento embutido no código-fonte de um software seja preservado, valorizado e compartilhado com todos adequadamente.”

A Recomendação sobre Ciência Aberta da Unesco destaca explicitamente a importância de software de código aberto como um componente-chave do conhecimento científico aberto, que claramente declara que:

“O conhecimento científico aberto se refere ao acesso aberto a publicações científicas, dados de pesquisa, metadados, recursos educacionais, software e código-fonte, e hardware, que estão disponíveis em domínio público, ou com copyright e licença aberta que permite



Foto: <https://dicosmo.org>. Reprodução

Figura 1. Roberto di Cosmo

acesso, reuso, reaproveitamento, adaptação e distribuição [...].”

O texto também enfatiza a necessidade de preservar e disponibilizar não apenas o código escrito por pesquisadores, mas também todas as suas dependências, cruciais para reprodutibilidade:

“No contexto de ciência aberta, quando o código-fonte aberto é um componente de um processo de pesquisa, para poder ser reusado e replicado geralmente é necessário que seja acompanhado de dados abertos

e especificações abertas do ambiente computacional necessário para compilá-lo e executá-lo.”

Isto está perfeitamente alinhado com a missão do Software Heritage de preservar todo o código-fonte publicamente disponível – incluindo suas dependências – desta forma dando apoio ao objetivo de fazer com que resultados de pesquisa sejam reprodutíveis, verificáveis e reusáveis com o passar do tempo (Figura 2).

C&C – Como podemos aumentar a percepção, entre cientistas, de que software é

um cidadão de primeira classe no ecossistema da Ciência Aberta?

RC – Para abraçar a Ciência Aberta em sua plenitude, a comunidade de pesquisa precisa reconhecer que o *software* é um resultado de pesquisa essencial, tanto quanto artigos e conjuntos de dados. Para fomentar este reconhecimento, há várias ações concretas que podem ser tomadas:

- Sistemáticamente, arquivar e preservar *software* de pesquisa para garantir que seja acessível e reproduzível. A Software Heritage tornou este processo extremamente simples, fornecendo extensões de navegadores que permitem arquivamento em um clique e integração via web com a maior parte das plataformas populares que hospedam código, como GitHub, GitLab, Gitea, ou Bitbucket.
- Citar o *software* de forma adequada usando identificadores intrínsecos e persistentes (chamados SWHIDs) que permitem referenciar precisamente vários níveis de um código (arquivos, revisões, *releases*). A Software Heritage recentemente inaugurou uma funcionalidade de “citação de *software*”, que produz informação exata de citação para todo o *software* que contenha os metadados adequados.
- Encorajar reconhecimento acadêmico de contribuições de *software*, estabelecendo prêmios para projetos de *software* de pesquisa exemplares. Esta iniciativa está sendo levada adiante de forma pioneira na França



Figura 2. Gráfico mostra que em janeiro de 2025 o arquivo do Software Heritage continha mais de 22 bilhões e meio de arquivos de código-fonte, quase 5 bilhões de “commits”, correspondendo a 345 milhões de projetos organizados em quase 18 bilhões de diretórios, 86 milhões de autores de software.

desde 2022, com um prêmio nacional de Ciência Aberta para *software*, descrito em detalhes no artigo de Blanc Catala et al.

- Integrar *software* em *workflows* de pesquisa para tornar seu uso, modificação e evolução rastreável.

O Software Heritage tem um papel-chave nesta transformação. Colabora com as principais infraestruturas acadêmicas, permitindo arquivamento e submissão de metadados de *software* de pesquisa de forma transparente. Editores e repositórios líderes como Zenodo, HAL, eLife, Episciences, e Dagstuhl já integram o Software Heritage a seus *workflows*, garantindo que o *software* de pesquisa seja preservado permanentemente e referenciado de forma adequada. Além disso, iniciativas

como ReplicabilityStamp.org for Computer Graphics aproveitam o Software Heritage para garantir reprodutibilidade computacional, reforçando o papel do *software* como um elemento fundamental em pesquisa. Ao embutir práticas de preservação e citação de *software* no ecossistema acadêmico, podemos elevar o *software* para torná-lo um resultado de pesquisa reconhecido e apreciado.

C&C – Como o Software Heritage contribuiu para avançar esta percepção?

Como instituições de pesquisa brasileiras podem participar desta grande iniciativa?

RC – O Software Heritage colabora ativamente com universidades, instituições de pesquisa e infraestruturas de todo o mundo para garantir que o *software* seja reconhecido,

arquivado e citado de forma adequada. As principais contribuições incluem:

- **Prover um sistema de arquivamento aberto:** O Software Heritage é um coletor proativo de código disponibilizado publicamente. Muitos pesquisadores já podem encontrar seu *software* preservado no arquivo, mesmo sem saberem que havia sido arquivado. O arquivamento garante que o *software* de pesquisa permaneça acessível sem exigir mecanismos complexos de registro.
- **Habilitar citação e reprodutibilidade de *software*:** O Identificador Hash de Software (SWHID) permite que pesquisadores possam referenciar seu código com identificadores persistentes, verificáveis e com padronização ISO. Esses

identificadores já estão sendo usados por editoras importantes, que assim reforçam seu papel na reprodutibilidade científica.

- **Integrar a infraestrutura-chave de pesquisa:** Plataformas como Zenodo e HAL agora sistematicamente transferem submissões de *software* para o Software Heritage, assim garantindo preservação e citação a longo prazo.
- **Lançamento do programa de adesão de ciência aberta:** instituições de pesquisa, bibliotecas e universidades podem apoiar ativamente o arquivo (do Software Heritage), ao mesmo tempo que obtêm acesso a treinamentos, oportunidades de colaboração e funcionalidades avançadas. As instituições brasileiras têm a oportunidade de aderir a esta rede, ajudando a moldar boas práticas em preservação de *software* de pesquisa.

C&C – Como as instituições de pesquisa brasileiras podem se envolver?

RC – As Universidades e os centros de pesquisa brasileiros podem contribuir para este movimento:

- **Garantindo que o *software* de pesquisa seja sistematicamente arquivado** no Software Heritage
- **Encorajando pesquisadores a usarem SWHIDs** para referenciar o seu *software*, melhorando seu reconhecimento e rastreabilidade.
- **Colaborando em iniciativas de Ciência Aberta**, integrando o Software Heritage as infraestruturas de pesquisa e programas de fomento nacionais.
- **Aderindo como membros ao programa de Ciência Aberta do Software Heritage**, que provê oportunidades de engajamento direto com uma rede global de instituições que trabalham com *software* de pesquisa, através do Grupo de Interesse de Arquivos e Bibliotecas.

Ao participar do Software Heritage, o Brasil pode liderar em Ciência Aberta e preservação digital do conhecimento, garantindo que o *software* de pesquisa produzido no país continue sendo um recurso valioso e acessível para gerações futuras. O Software Heritage não é apenas um arquivo – é uma infraestrutura global para Ciência Aberta. Sua integração com grandes editoras, repositórios de pesquisa e iniciativas

acadêmicas faz dele uma ferramenta essencial para preservar, referenciar e fazer com que *software* de pesquisa seja reconhecido. Pesquisadores e instituições brasileiros agora têm uma oportunidade única para aproveitar o Software Heritage para fortalecer a Ciência Aberta e assegurar que o *software* permaneça como um componente confiável e acessível do conhecimento científico.

Roberto di Cosmo é professor de Informática na Universidade de Paris Diderot e pesquisador no Inria. Especialista em *software* livre e ciência aberta, fundou o Software Heritage e liderou iniciativas como o projeto Mancoosi e o IRILL. Preside o IMDEA Software e o Comitê Nacional para a Ciência Aberta na França.

Fernanda Antonia da Fonseca Sobral é professora emérita da UnB e pesquisadora emérita do CNPQ, colaborando com o Programa de Pós-Graduação em Sociologia (UnB). Foi Vice-Presidente da SBPC em duas gestões (2019-2021 e 2021-2023) e atualmente é Diretora da SBPC.

Claudia Bauzer Medeiros é professora do Instituto de Computação da Universidade Estadual de Campinas (Unicamp), além de membro da Academia Brasileira de Ciências (ABC) e Fellow da World Academy of Sciences (TWAS).