Comunicação

DOI:10.21680/2447-7842.2023v9n2ID33759

dARK: uma implementação descentralizada de identificadores persistentes ARK baseada em blockchain

dARK: a decentralized blockchain implementation of ARK Persistent Identifiers

Washington Luís Ribeiro de Carvalho Segundo [1]
Lautaro Matas [2]
Thiago Nóbrega [3]
José Edilson S. Filho [4]
Jesús P. Mena-Chalco [5]

Submetido em: 17/04/2023	Aprovado na ConfOA: 14/06/2023	Publicado em: 25/11/2023

Resumo: Apresentamos a primeira prova de conceito de uma tecnologia que pode ser a base para um serviço descentralizado e de baixo custo para atribuir/resolver identificadores ARK persistentes. Esta é a base para um projeto aberto e voltado para a comunidade promovido pelo IBICT (Brasil) e LA Referencia/RedCLARA com o objetivo de, a longo prazo, fornecer uma fábrica aberta/não centralizada de identificadores persistentes únicos/deduplicados e um serviço de resolução para o ecossistema global de Ciência Aberta baseado na tecnologia blockchain pública permissionada. Algumas das motivações para este trabalho são: (i) A necessidade não apenas de identificadores persistentes, mas únicos/deduplicados, a fim de construir melhores grafos de pesquisa, indicadores e dados de avaliação da

[1] Doutor em Informática.
[2] Especialista em Ciência da Computação.
[3] Doutor em Ciência da Computação.
[4] Mestre em Engenharia de Teleinformática.
[5] Doutor em Ciências da Computação.

pesquisa; (ii) Falta de cobertura do PIDs em repositórios do Sul Global principalmente pelos custos que esses serviços representam para as instituições; e (iii) A maioria dos sistemas de identificadores persistentes são baseados em modelos centralizados, dependendo de poucas agências que suportam a infra-estrutura de serviços. O identificador ARK surgiu como uma alternativa viável de solução de baixo custo devido à possibilidade de implementação de provedores internos para o resolvedor global. dARK é uma implementação ARK baseada em nós de blockchain institucionais, os dados são de propriedade, armazenados e controlados de forma distribuída por todas as organizações mantenedoras da rede, sem a necessidade de intenso uso de recursos computacionais.

Palavras-chave: Archival resource key - ARK; identificadores persistentes; blockchain.

Abstract: We present the first proof of concept of a technology that could be the basis for a decentralized, low-cost service for assigning/resolving persistent ARK identifiers. This is the basis for an open, community-driven project promoted by IBICT (Brazil) and LA Referencia/RedCLARA with the long-term goal of providing an open/non-centralized persistent unique/duplicate identifier factory and resolution service for the global Open Science ecosystem based on permissioned public blockchain technology. Some of the motivations for this work are: (i) The need for not only persistent, but unique/deduplicated identifiers in order to build better research graphs, indicators and research evaluation data; (ii) Lack of coverage of PIDs in repositories of the Global South mainly due to the costs that these services represent for institutions; and (iii) Most persistent identifier systems are based on centralized models, depending on a few agencies that support the service infrastructure. The ARK identifier has emerged as a viable alternative low-cost solution due to the possibility of implementing internal providers for the global resolver. dARK is an ARK implementation based on institutional blockchain nodes, the data is owned, stored and controlled in a distributed manner by all the organizations maintaining the network, without the need for intensive use of computing resources.

Keywords: Archival resource key - ARK; persistent identifiers; blockchain.

1 INTRODUÇÃO

No contexto científico, os Identificadores Persistentes (PIDs) são essenciais em vários aspectos, como por exemplo: citação; crédito e autoria; pedidos de patente; proveniência do conhecimento; e validação (McMurry *et al.*, 2017; Figueiredo, 2017; *Sansone et al.,* 2019. No atual ecossistema de pesquisa, os PIDs são a chave para identificar artigos, livros, autores, documentos, arquivos, bancos de dados, amostras, objetos de arte e páginas, entre outros objetos científicos digitais, em todo o mundo aberto nacional, regional e global em ecossistemas científicos (McMurry *et al.*, 2017).

Os serviços atuais de Identificadores Persistentes (PIDs) dependem principalmente de organizações e comunidades sem fins lucrativos. Os dados são armazenados e controlados por um pequeno número de agências que fornecem a atribuição e resolução dos PIDs, além de implementar serviços para organizações de pesquisa. As organizações de pesquisa geralmente contribuem financeiramente para sustentar os serviços e a infraestrutura necessária. Os custos operacionais e de manutenção são cobertos por meio de diferentes modelos, como taxas anuais, associações, taxas de criação de PIDs, entre outros. Por vezes, esses custos representam uma barreira para instituições pequenas, especialmente em regiões menos desenvolvidas. Como resultado, há uma falta de cobertura de PIDs em muitas partes do mundo.

Uma alternativa para atribuir e resolver PIDs são os identificadores ARK (Archival resource key). O modelo ARK será abordado adiante, mas resumidamente, pode-se dizer que o processo de atribuição deste identificador é gerenciado de maneira autossuficiente. Não é necessária uma autoridade central para gerar um novo identificador, principalmente porque o ARK não precisa de informações externas para atribuir um novo identificador. Portanto, algumas instituições implementaram soluções ARK internas, o que pode ser mais acessível para algumas instituições, mas traz outros problemas e limitações, como riscos de perda de dados.

Archival Resource Key (ARK) é um sistema PID aberto que fornece referências confiáveis para objetos de informação. Segundo Kunze e Bermès (2008) na obra “The ARK identifier scheme”, Archival Resource Keys (ARKs) são uma alternativa flexível e de baixo custo para atribuição de identificadores persistentes. Qualquer organização tem a possibilidade de criar uma quantidade ilimitada de identificadores usando um esquema de metadados flexível. No entanto, a maioria das implementações ARK depende de soluções internas que são isoladas umas das outras, trazendo alguns problemas como duplicação de PID, ineficiência de custo e baixa tolerância a falhas.

Apresentamos a primeira implementação ARK (prova de conceito) de um serviço PID descentralizado aberto concebido desde o início como um bem público para o ecossistema de Ciência Aberta. A tecnologia por trás, blockchain, permite que a governança e os baixos custos sejam compartilhados por redes nacionais e regionais de pesquisa/educação em colaboração com institutos/universidades de pesquisa individuais, com o fornecimento de pequenos recursos computacionais.

Propomos executar o ARK em cada nó de uma *Blockchain Consortium Network* (BCN) implementando o dARK como *Blockchain Decentralized Application* (DApp). Esse layout de arquitetura mitiga problemas relacionados ao acesso contínuo ao sistema PID ao longo do tempo e cria mecanismos padrão para integrar PIDs de diferentes fontes a baixo custo para as organizações que aderem à BCN. Como um benefício colateral do uso da tecnologia blockchain, o dARK fornece proveniência PID nativa, adicionando uma nova camada de confiança aos metadados do ARK. No entanto, no presente artigo, omitimos os detalhes técnicos que podem ser encontrados em Washington Segundo *et al.* (2022).

2 ARCHIVAL RESOURCE KEY (ARK)

As Archival Resource Keys (ARKs) são um sistema aberto de PID que fornece referências confiáveis para objetos de informação. As ARKs são amplamente utilizadas como sistema de PID e foram adotadas por mais de 900 organizações. Essas organizações usam as ARKs para gerar mais de 8 bilhões de

identificadores, desde objetos digitais (como documentos e bancos de dados) até objetos físicos (amostras biológicas e obras de arte), e até mesmo seres vivos (como pessoas e orquestras) e objetos intangíveis (lugares e termos de vocabulário) (Kunze, 2021*)*.

Para serem usadas nas várias aplicações mencionadas acima, as ARKs foram concebidas para serem genéricas. Além disso, os princípios fundamentais do design das ARKs são:

1. Identificadores de alta densidade, permitindo identificadores opacos e com comprimento reduzido (*short link*);
2. Autossuficiência, um servidor simples local consegue executar as ferramentas necessárias para se gerar ARKs;
3. Esquema de metadados flexível;
4. É Acessível: não há taxas a serem pagas para se atribuir ARKs.

Deve-se notar que a autossuficiência do identificador ARK é ideal para sistemas descentralizados. Essa característica, combinada com a ausência de taxas, esquema de metadados flexível e a natureza de código aberto da ARK, são os requisitos que combinam perfeitamente com a solução desejada de um PID plural e inclusivo.

3 BLOCKCHAIN

A maioria dos sistemas de computador e aplicativos considera que uma autoridade central deve controlar os dados e/ou funções dos sistemas. Nesse caso, os sistemas e aplicativos pressupõem um sistema de gerenciamento centralizado. Por exemplo, organizações financeiras (como bancos) e até mesmo o Google Search são considerados centralizados porque há uma organização no centro que administra todo o negócio, incluindo dados, designers, programadores e especialistas em publicidade.

Além disso, um sistema centralizado requer um sistema de backup e alta disponibilidade complexo para lidar com a desvantagem do ponto único de falha. Finalmente, vale a pena comentar sobre a existência de falhas tecnológicas; por exemplo, se uma revista científica é descontinuada, todo o histórico de publicações da revista pode ser perdido. Esse tipo de falha também pode ocorrer em sistemas de identificadores persistentes.

Sistemas descentralizados surgiram para mitigar as limitações dos sistemas centralizados. Em vez de usar uma autoridade central para executar e orquestrar operações, essas operações são realizadas pelos nós de uma rede descentralizada. Nesse cenário, o controle do sistema é igualmente compartilhado entre os nós da rede, usando protocolos criptográficos e de segurança.

Blockchain, também chamado de registro distribuído, é essencialmente um sistema de gerenciamento de banco de dados mantido por um conjunto de nós que não confiam plenamente um no outro. *Blockchain* é uma tecnologia que mantém os estados e as transações históricas, sem nenhum nó central para impor conformidade. A *blockchain* fornece armazenamento imutável (garantia de evidência de violação) das transações em uma cadeia de blocos, armazenando dados (registros) em blocos que são vinculados usando ferramentas criptográficas.

Ao se considerar um sistema descentralizado que utiliza a tecnologia *blockchain*, tem-se as seguintes vantagens:

• Segurança aprimorada: Devido às ferramentas criptográficas robustas empregadas no núcleo desses sistemas, cada usuário deve ser identificado e assinar cada transação usando suas chaves criptográficas. Essa característica adiciona uma camada de segurança aos sistemas;

• Transparência e confiabilidade: Cada transação realizada nesse sistema é verificada por todos os nós da rede. Todas as transações e dados armazenados nesses sistemas são auditáveis, tornando o sistema transparente e confiável;

• Garantias de evidência de violação: Os dados armazenados não podem ser modificados após inserção. Contrapondo um sistema centralizado, onde qualquer pessoa no topo da hierarquia com autorização pode fazer alterações;

• Procedência de dados inseridos: uma vez que os dados são imutáveis e cada transação indica quem e quando foi executada, esses sistemas têm um sistema incorporado de procedência de dados.

Essas vantagens podem ser usadas para melhorar os sistemas de identificadores persistentes existentes. Por exemplo, a característica de transparência de dados pode ser usada para promover mecanismos de integração de identificadores persistentes, e o benefício da procedência de dados pode ser utilizado no contexto de identificadores persistentes.

4 IDENTIFICADORES DESCENTRALIZADOS DARK

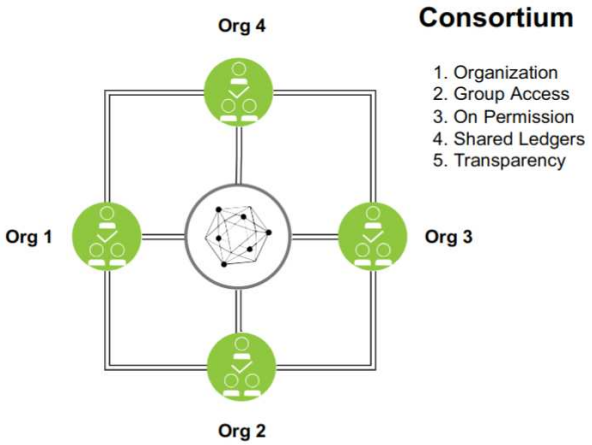
O sistema de atribuição de identificadores ARK descentralizados, que chamamos de dARK, emprega uma abordagem que permite que várias instituições gerenciem (colaborativamente) seu sistema de identificadores persistente ARK, usando uma infraestrutura descentralizada comum, baseada em nós de uma rede (consórcio) *blockchain*. Dessa forma, os dados não são de propriedade, armazenamento ou controle de uma única organização, mas de todos os participantes da rede.

Considerando que os usuários do dARK participarão de uma rede *blockchain* de consórcio bem organizada, os dados do PID serão replicados de maneira segura, confiável, auditável e acessível em todos os nós dessa rede. Mesmo que a instituição desapareça, os dados armazenados serão persistidos nos nós restantes da rede. Além disso, considerando a rede *blockchain* do consórcio, o custo de armazenar e gerenciar os dados do PID será baixo. Estima-se que para cada atribuição de identificador dARK, são considerados 12KB de armazenamento em disco. Além disso, foi projetada uma instância padrão de servidor para executar o aplicativo descentralizado (por exemplo, no ambiente experimental, foi usada uma máquina virtual com 4 núcleos Xeon, 16GB de RAM e 30GB de armazenamento).

Um consórcio *blockchain* é um tipo de rede semi-descentralizada de estilo de rede permissionada, onde nós ou membros se juntam à rede por meio de uma entidade reguladora. Em geral, esse sistema é baseado em votação para garantir

baixa latência e altas taxas de velocidade. Cada nó tem permissão para gravar transações, mas não pode adicionar um bloco por si só e cada bloco adicionado por outro nó precisa ser verificado antes de ser adicionado à rede. A Figura 1 representa um exemplo de consórcio formado por 4 organizações. Todos os nós da rede se comunicam entre si. Em um eventual número par de organizações, outros algoritmos de consenso diferentes da votação por maioria simples podem ser escolhidos.

Figura 1 - Consórcio Blockchain



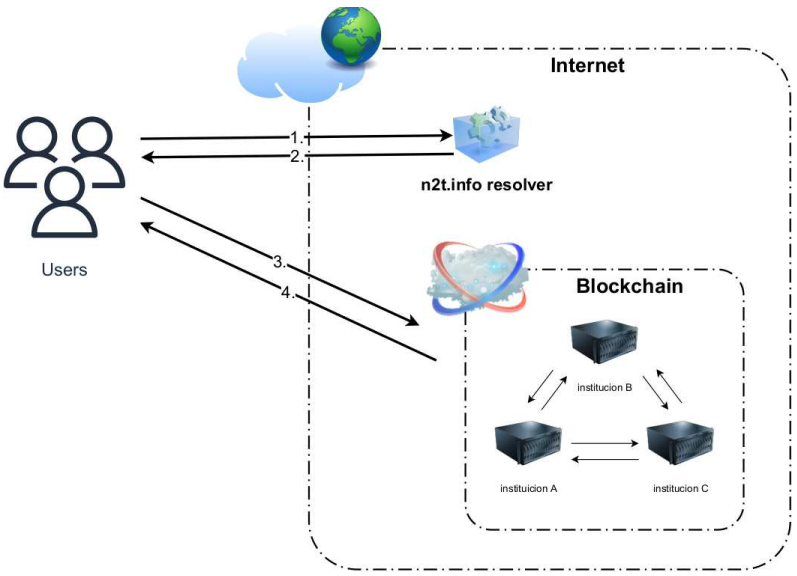
Fonte: Os autores.

5 PROVA DE CONCEITO DO DARK

O sistema dARK pode ser descrito como um aplicativo cliente-servidor, onde o servidor é a rede de consórcio blockchain dARK e o cliente pode ser implementado em várias tecnologias (por exemplo, páginas da web HTML, repositórios ou sistemas de revistas eletrônicas). A Figura 3 ilustra a arquitetura implementada na PoC. Foi usada a tecnologia baseada em *Ethereum*, chamada *Hyperledger Besu*, como a rede de consórcio blockchain. Além disso, a opção por essa tecnologia foi feita com base em algumas vantagens:

1. Tecnologia de código aberto amplamente usada (suportada pela *Linux Foundation*);
2. Capaz de criar uma rede de consórcio blo ckchain privada;
3. Compatível com várias redes de *blockchain* abertas (por exemplo, a *Ethereum mainnet*);
4. Possui bibliotecas nativas para várias linguagens de programação (por exemplo, PHP, Java, Python, Go, C/C++, JavaScript);

Figura 2 - Serviço dARK



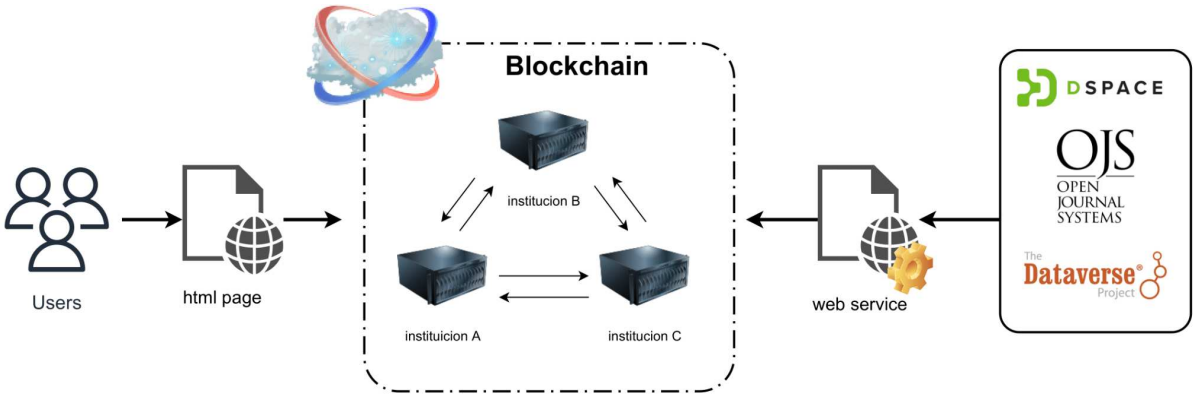
Fonte: Os autores.

Em relação ao lado do cliente do dARK, a Figura 3 ilustra dois tipos diferentes de cliente. No lado esquerdo da figura, há um exemplo em que os curadores (usuários) atuam diretamente em seus navegadores da web para interagir com o sistema dARK. É importante mencionar que a página é uma aplicação simples de HTML+JavaScript. Além disso, uma vez que cada operação é realizada pela *blockchain* (atuando como um servidor web), a página HTML é usada para testar a interação com o sistema dARK. No lado direito da Figura 3, é mostrado um serviço da web que pode integrar o sistema dARK a software existente, como DSpace,

Open Journal System (OJS) ou Dataverse. As ações de interoperabilidade via serviços da web são as mesmas do teste da página HTML. No entanto, é necessário implementar um plug-in do lado do cliente (DSpace e outros), para gerar a configuração da carteira de blockchain. O plug-in pode ser um trecho de código *JavaScrip*t desenvolvido para diferentes plataformas de cliente, que pode ser facilmente adicionado ao pipeline de registro local. Depois de registrar um objeto de publicação, o registro armazenado possui os seguintes atributos:

1. *noid*: é a cadeia de sufixo do identificador dARK;
2. *external_pids*: Uma matriz que contém informações sobre os outros identificadores mapeados que já foram já foram atribuídos ao registro do objeto;
3. *payload*: A carga útil do registro de objeto (descrição sucinta do recurso);
4. *link*: URL de link externo que aponta para o objeto registrado;
5. *owner*: o endereço da carteira do registrante do registro de objeto.

Figura 3 - Prova de conceito dARK



Fonte: Os autores

6 CONSIDERAÇÕES FINAIS

Vemos este trabalho como a base de uma fábrica de PIDs abertos/não centralizados/ e deduplicados, além de serviço de resolução de entidades para o ecossistema global de Ciência Aberta, com base na tecnologia blockchain pública permissionada. É importante salientar que a implementação proposta não exige intenso uso de recursos computacionais. A curto prazo, esperamos fornecer uma solução PID para o sul global, onde apenas uma pequena parte das instituições tem recursos para atribuir PIDs. Os trabalhos em curso, que se espera serem apresentados na conferência, incluem melhorias na integração entre os sistemas dARK e ARK, nova implementação com melhor aproveitamento das chaves de pesquisa, recuperação de registos com consultas mais avançadas aos metadados PID, mas mantendo a simplicidade e respostas rápidas da tecnologia blockchain. Além disso, planejamos avançar nas implementações de clientes para o dARK, principalmente em plataformas de código aberto, como DSpace, OJS e Dataverse.

REFERÊNCIAS

Figueiredo, A. S. (2017). Data sharing: convert challenges into opportunities. *Frontiers in public health*, 5, e327. DOI: 10.3389/fpubh.2017.00327

Kunze, J., and Bermès, E. (2008). “The ARK identifier scheme” .

Kunze, J. (2021). *ARK Alliance: Empowering 800 institutions and 8 billion identifiers since 2001*. In iPRES.

McMurry, J. A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., Courtot, M., Deck, J., Dumontier, M., Fellows, D. K., Gonzalez-Beltran, A., Gormanns, P.,

Grethe, J., Hastings, J., Hériché, J.-K., Hermjakob, H., Ison, J. C., Jimenez, R. C., Jupp, S., ... & Parkinson, H. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLOS Biology*, *15*(6), e2001414. https://doi.org/10.1371/journal.pbio.2001414

Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L., & Thurston, M. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, *37*(4), 358–367. https://doi.org/10.1038/s41587-019-0080-8

Washington Segundo, W., Matas, L., Nóbrega, T., S., J. E., Filho, & Mena-Chalco, J. (2022). *dARK: A decentralized blockchain implementation of ARK Persistent Identifiers*. https://doi.org/10.5281/zenodo.7686101