Cariniana
Rede Brasileira de Serviços de Preservação Digital

ibict 60 anos

Ministério da Ciência, Tecnologia e Inovação

GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA

Relatório Técnico
Análise das Caixas LOCKSS

Arthur Heleno Lima R. de Souza

Agosto - 2014

Identificação do Documento	
Projeto	Pesquisa e desenvolvimento de serviços de preservação digital em rede
Rede	Cariniana – PLN/SEER
Identificação do Relatório	RTA001
Período de coleta de dados	Março /2013 – Abril /2014
Autor	Arthur Heleno Lima R. de Souza
Coordenador	Miguel Ángel Márdero Arellano
Resultado	Relatório técnico contendo a análise das informações obtidas nas caixas LOCKSS IBICT e parceiros.

Sumário

1. Introdução

A Rede CARINIANA surgiu da necessidade de se preservar documentos eletrônicos brasileiros a fim de garantir o acesso a longo prazo dos conteúdos armazenados digitalmente. A infraestrutura tecnológica da Rede segue o modelo de uma PLN (Rede Privada LOCKSS), no qual existe uma conexão entre "nós" de uma rede e cada nó possui o software LOCKSS sendo executado para coletar os dados a serem preservados. Este software instalado em cada nó da rede também é chamado de caixa LOCKSS, e cada caixa deverá ter o mesmo conteúdo de uma outra caixa.

A PLN da Rede Cariniana preserva periódicos eletrônicos cadastrados no portal SEER ou Sistema Eletrônico de Editoração de Revistas, e a alimentação da rede procede de maneira periódica, a fim de possibilitar a analise dos dados.

Os procedimentos desta análise de dados tiveram seu momento inicial em 2013, quando o IBICT/Rede Cariniana já havia estabelecido vínculo de parceria com cinco instituições de pesquisa: Universidade de São Paulo, Universidade Federal de Santa Maria, Universidade Estadual de Campinas, Universidade Estadual do Maranhão, Universidade Federal da Paraíba.

O IBICT, respaldado pela equipe do programa LOCKSS em Stanford, e cada uma das instituições parceiras citadas anteriormente instalou uma caixa LOCKSS, formando a PLN: Rede Privada LOCKSS de periódicos eletrônicos.

2. Objetivos

- Disponibilizar informações referentes aos processos de coleta e preservação das caixas LOCKSS;
- Analisar e interpretar os erros e avisos disparados nas caixas;
- Estudar e apresentar síntese dos dados obtidos;

3. Procedimentos

3.1. Obtenção de dados

Os dados para análise foram obtidos através de três principais fontes:

a) Interface Administrativa LOCKSS
b) Acesso SSH às caixas LOCKSS
c) Contato com a Equipe LOCKSS da *Stanford University*

A Interface Administrativa é uma ferramenta do sistema LOCKSS capaz de apresentar dados sobre os principais processos de preservação, como o rastreio, coleta, armazenamento, teste de integridade e etc. Também possibilita o administrador a interagir com o sistema, com a possibilidade de mudar parâmetros, adicionar novo conteúdo, modificar a fila de ações e outras funcionalidades.

Nesta Interface foi possível obter a maior parte dos dados analisados, uma vez que esta permite visualizar, em tempo real, os processos orientados a periódicos. Acessar, via SSH, o servidor onde o software LOCKSS está em execução possibilita verificar o estado da máquina, portas da rede e problemas que possam surgir no eventual dos processos em execução.

O contato que a equipe do IBICT da Rede Cariniana manteve com a equipe LOCKSS foi importante para manter os serviços da rede ativos. A experiência e o conhecimento transmitido possibilitou sanar muitos problemas encontrados, manter os processos de preservação e dar andamento com o desenvolvimento de novas ferramentas.

Os dados obtidos foram levantados em seis caixas LOCKSS, cada caixa lotada em uma das seis instituições parceiras do projeto piloto da Rede Cariniana: Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal de Santa Maria, Universidade Estadual de Campinas, Universidade de São Paulo, Universidade Federal da Paraíba e Universidade Estadual do Maranhão.
Cada instituição na rede pode ser referida como um “nó”.

3.2. Análise de dados

A análise de dados é um processo de inspeção, transformação e verificação com o objetivo de descobrir informações úteis, sugerindo conclusões, e apoiar a tomada de decisão. Este processo teve o seu início em março de 2013, quando os primeiros testes do sistema LOCKSS na rede Cariniana foram executados.

3.2.1. Periódicos

O primeiro teste com preservação de periódicos na Rede Cariniana ocorreu na adição de 15 volumes para o teste da PLN, 6 da revista Ciência da Informação (IBICT) e 9 da Zetetike: Revista de Educação Matemática (Unicamp). O funcionamento das caixas ocorreu sem erros, efetuando os processos de coleta (*harvest*) de dados, testes de erros, integridade, verificação e, consequentemente, obteve-se a preservação das revistas.

Após este processo, as equipes continuaram a tratar mais periódicos para que estes pudessem ingressar na rede. Porém, um obstáculo foi observado: periódicos sem padronização, faltando arquivos ou com erros em sua configuração. Para contornar esta situação, o IBICT e a *Stanford University* uniram esforços para tratar os objetos digitais irregulares de acordo com as etapas de preservação adotados pela rede LOCKSS.

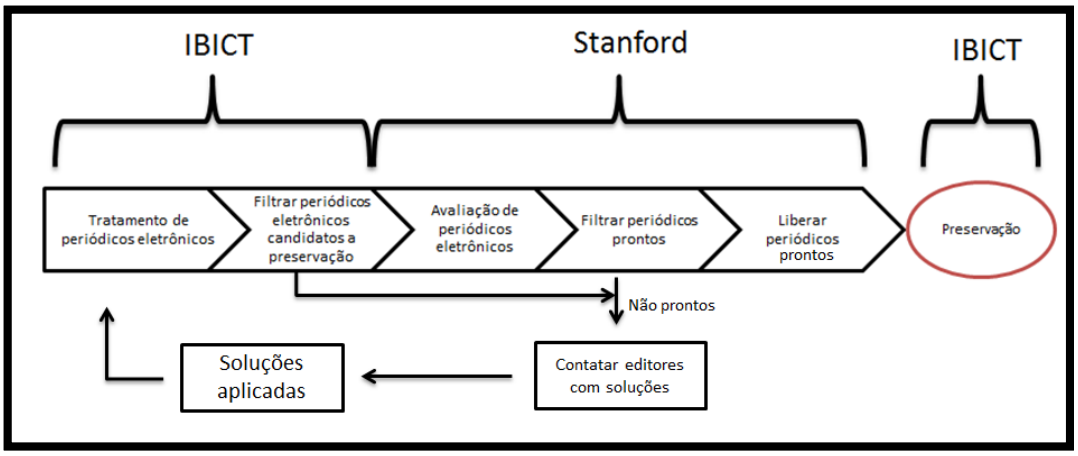


Fig. 1 – Fluxo de tratamento de periódicos

Observa-se na Fig.1, Tratamento de periódicos eletrônicos, que o periódico é analisado pelo IBICT e procuram-se problemas. Caso existam erros, são aplicadas

as soluções cabíveis. Desta forma, no segundo passo, selecionam-se os periódicos candidatos envia-os para a Equipe em Stanford, enquanto os periódicos não selecionados são reanalisados.

No terceiro e quarto passo, os periódicos aprovados pelo IBICT irão passar pelo crivo da equipe de *Stanford*. Serão avaliados mais uma vez e filtrados, os não-prontos serão reportados de volta ao IBICT e os prontos irão ser coletados para a preservação. Os não-prontos, ou não selecionados, possivelmente possuem erros que só podem ser solucionados pelos detentores do periódico. Uma proposta de solução será enviada para os responsáveis.

	Títulos	Volumes
Preservação	281	2008
Tratamento	621	6434
Total	902	8442

Tabela 1 – Periódicos da Rede Cariniana

Em abril de 2014, a Rede Cariniana contabilizou 6434 volumes de 621 títulos de periódicos eletrônicos no processo de tratamento, e 2008 volumes de 281 títulos já preservados na rede, totalizando 8442 volumes de 902 títulos. Paralelamente, mais periódicos estão sendo selecionados para o tratamento de preservação, incluindo solicitações de instituições que desejam ter seus periódicos na rede.

3.2.2. Estado da Caixas

Como um exemplo de implantação do sistema, é possível citar as especificações da caixa LOCKSS IBICT, que está em execução em uma maquina virtual de um servidor:

- Uma CPU de um núcleo, 2 GHz
- 2 GB de memória
- Linux CentOS 5.1
- 500 GB em disco

No caso do armazenamento, este critério depende da abordagem da iniciativa de preservação digital, como visto no **subitem 3.2.2**. As especificações de processamento e memória das outras caixas de outras instituições não puderam ser obtidas no intervalo da obtenção de dados.

A distribuição mais comum do LOCKSS é através de uma imagem de disco contendo o sistema operacional LINUX CentOS e o software propriamente dito. Porém, é possível instalar o software de outras formas e em outras distribuições LINUX, como é o caso da Universidade Federal da Paraíba que realizou a implantação do sistema no Debian 6. Um relato sobre o *"know-how"* será disponibilizado pela UFPB.

Os dados da infraestrutura das caixas LOCKSS:

Instituição	IP	Estado	Espaço em disco disponível	Espaço em disco utilizado	UAs coletadas	OS - LINUX
IBICT	200.130.0.165	Em execução	457 GB	9.5 GB	502	CentOS
UFSM	200.18.45.26	Em execução	62 GB	9.4 GB	502	CentOS
Unicamp	143.106.108.25	Em execução	887 GB	9.4 GB	502	CentOS
USP	143.107.154.11	Desconectada	(?)	(?)	(?)	CentOS
UFPB	150.165.241.6	Em execução	50 GB	7.8 GB	502	Debian 6
UEMA	200.137.128.139	Em execução	81 GB	9.3 GB	502	CentOS

Tabela 2 – Dados da infraestrutura: 29 de outubro de 2013

Os valores da **Tabela 2** foram extraídos ao final de outubro, quase um mês após a quinta adição de periódicos, que ocorreu no dia 2 deste mesmo mês. Nesse período as caixas LOCKSS já estavam sincronizadas e com testes de auditorias e integridade em dia, desta forma foi possível obter valores estáveis para análise, porém, a caixa USP estava desconectada devido à ocupação da reitoria da universidade ocorrida neste período.

É possível verificar que o uso em disco das caixas é semelhante, com exceção da UFPB – Fenômeno pode ser explicado pela utilização de outra distribuição Linux, contudo não afeta em suas funcionalidades.

Instituição	IP	Estado	Espaço em disco disponível	Espaço em disco utilizado	UAs coletadas	OS
IBICT	200.130.0.34	Em execução	457 GB	46 GB	2008	CentOS
UFSM	200.18.45.26	Em execução	62 GB	46 GB	2008	CentOS
Unicamp	143.106.108.25	Em execução	887 GB	41 GB	2008	CentOS
USP	200.144.183.66	Em execução	4 TB	52 GB	2008	CentOS
UFPB	150.165.241.6	Em execução	50 GB	42 GB	2008	Debian 6
UEMA	200.137.128.139	Em execução	81 GB	47 GB	2008	CentOS

Tabela 3 – Dados da infraestrutura: 20 de janeiro de 2014 de 2014

O momento da extração de dados da **Tabela 3** ocorreu poucos dias após a sétima adição de volumes de periódicos, sendo perceptível a diferença do espaço utilizado em cada caixa. Isso ocorre, pois o *harvest* e os testes de integridade acontecem de forma assíncrona.

3.2.3. Comunicação das Caixas LOCKSS

Os nós da rede Cariniana se comunicam através do protocolo P2P e, diferentemente do diagrama cliente-servidor, as mensagens são enviadas em uma hierarquia de rede planificada, cada nó enviando e recebendo dados. Essa comunicação entre as caixas é de suma importância para que os processos de verificação e de sincronia sejam efetuados.

Desta forma, é necessário que os administradores da rede em que o sistema LOCKSS está em execução preparem o ambiente para que não impeça o funcionamento correto e ainda mantenha a segurança adequada.

Para configurar o ambiente, torna-se necessário configurar certas portas para alguns destinos, como explicitado na tabela abaixo.

Porta	Descrição	Destino
22	SSH	IBICT Stanford
8080	Servidor de conteúdo/PROXY	IBICT Stanford
8081	Interface Administrativa	IBICT Stanford Administrador
9729	V3 LCAP	Caixas

Tabela 4 - Conexões de entrada (*inbound*)

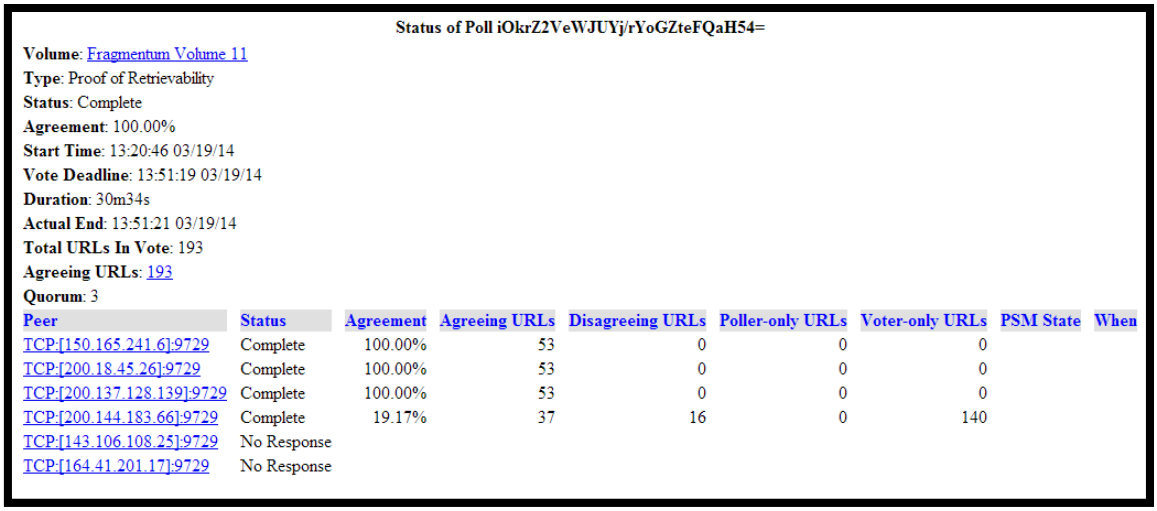
A Tabela 4 mostra as portas utilizadas pelo LOCKSS. Embora o funcionamento autônomo das caixas necessite apenas da porta 9729, a gerência e administração requerem as outras citadas acima. A porta 22, com acesso a máquina será acessado pela equipe IBICT e Stanford caso seja reportado problemas ou erros, ou para realizar configurações que algum parceiro esteja com dificuldades.

A porta 8080 de acesso ao Servidor de conteúdo destinara-se ao IBICT e Stanford por questões de monitoramento e dados estatísticos.

A 8081 hospeda uma página web com as ferramentas administrativas do sistema LOCKSS, destinado para o administrador de cada caixa LOCKSS e para IBICT e Stanford, a fim de auxiliar nos processos de preservação.

A porta 9729 utiliza um serviço da V3 LCAP, um protocolo de comunicação ponto a ponto, baseado em P2P, que é fundamental para que o software consiga se comunicar com as outras caixas em uma PLN. Essa porta também poderá ser liberada para todas as conexões caso a política da instituição permita, ou criar todas as exceções para os endereços IP de cada caixa LOCKSS.

Um dos processos fundamentais realizados nesta porta é o teste de integridade via o processo de “votes” e “polls”. Esses processos comparam os dados entre as caixas para auditar o conteúdo e reparar os objetos digitais no caso de perdas ou danos.



Status of Poll iOkrZ2VeWJUYj/rYoGZteFQaH54=

Volume: Fragmentum Volume 11
Type: Proof of Retrievability
Status: Complete
Agreement: 100.00%
Start Time: 13:20:46 03/19/14
Vote Deadline: 13:51:19 03/19/14
Duration: 30m34s
Actual End: 13:51:21 03/19/14
Total URLs In Vote: 193
Agreeing URLs: 193
Quorum: 3

Peer	Status	Agreement	Agreeing URLs	Disagreeing URLs	Poller-only URLs	Voter-only URLs	PSM State	When
TCP:[150.165.241.6]:9729	Complete	100.00%	53	0	0	0		
TCP:[200.18.45.26]:9729	Complete	100.00%	53	0	0	0		
TCP:[200.137.128.139]:9729	Complete	100.00%	53	0	0	0		
TCP:[200.144.183.66]:9729	Complete	19.17%	37	16	0	140		
TCP:[143.106.108.25]:9729	No Response							
TCP:[164.41.201.17]:9729	No Response							

Fig.2 Resultado do processo de *Poll*

Analisando a Fig.2, pode ser visto os resultados de um processo de verificação de integridade. Duas caixas não participaram: uma por estar com dificuldades técnicas e se apresentar off-line no momento em que foi extraído este dado, e a outra por não ter configurado sua caixa para se comunicar com as outras pela porta 9729. Cinco caixas LOCKSS participaram do processo: a que iniciou o processo, IBICT (200.130.0.34), e outras quatro que possuem o “Complete” na segunda coluna da tabela da figura.

Desse processo, entende-se que os dados coletados pela caixa IBICT foram idênticos aos coletados por três outras caixas. Porém, uma das caixas participantes obteve um índice baixo de concordância. Infere-se, através do processo de integridade, que a caixa USP (200.144.183.66) possui inconsistências nos dados coletados da revista eletrônica Fragmentum Volume 11, e por consequência, irá reparar esses dados.

Esse processo acontece diversas vezes e para cada objeto preservado. O objetivo é garantir a integridade do conteúdo em preservação.

3.2.4. Previsão de crescimento

Como visto anteriormente, a Rede Cariniana passou por um aumento significativo de periódicos incluídos em seu processo de preservação, desta forma, torna-se indispensável analisar esse aumento para prever o futuro da rede. Os dados coletados na Tabela 2 deram subsídios para o cálculo.

Adição	Volumes	Armazenamento em disco	Armazenamento /Volumes
1	15	345 MB	23 MB
2	82	1558 MB	19 MB
3	51	969 MB	19 MB
4	43	860 MB	25 MB
5	311	7151 MB	24 MB
6	593	13046 MB	24 MB
7	913	20999 MB	23 MB
TOTAL	2008	~45 GB	MÉDIA: ~22,5 MB

Tabela 2 – Dados do armazenamento

A prospecção destes dados baseou-se na coleta do armazenamento médio entre as caixas para cada adição de volumes. Este valor foi dividido pelo o número de volumes preservados, resultado em um valor médio, e deste valor médio de cada adição geramos outra média no valor de 22,5 Megabytes.

Valores	Média	Desvio	Quadrado dos desvios
23	22,5	0,5	0,25
19	22,5	-3,5	12,25
19	22,5	-3,5	12,25
25	22,5	2,5	6,25
24	22,5	1,5	2,25
24	22,5	1,5	2,25
23	22,5	0,5	0,25
Soma dos quadrados dos desvios:			35,75

Tabela 3 – Desvios

Na Tabela 3, observa-se o desvio e o quadrado do desvio das ocorrências. Com a soma destes dados, pode-se calcular a variância da média:

$$V = \frac{35{,}75}{7} = \sim 5{,}1$$

A soma dos quadrados dos desvios dividida pelo número de ocorrências resulta na variância da média anteriormente calculada. Esse valor pode ser usado para calcular o desvio padrão:

$$Dp = \sqrt{V} = \sqrt{5{,}1} = \sim 2{,}26$$

O valor do desvio padrão é baixo, o que significa que os dados analisados são homogêneos e confiáveis para o objetivo do calculo: prever o crescimento da rede orientado ao ingresso de periódicos eletrônicos.

A rede está utilizando aproximadamente 45 Gigabytes de armazenamento em cada caixa para preservar 281 títulos (2008 volumes). Em teoria, para preservar todos os 8442 volumes de 902 títulos, a rede utilizaria:

$$C = 8442\ x\ 22{,}5 = \sim 189945$$

É possível prognosticar que o uso em disco da rede seria cerca de 190 Gigabytes para armazenar a totalidade em volumes em tratamento e as já preservadas.

Visto os dados levantados, torna-se necessário alocar em disco uma quantidade superior à calculada, uma vez que sistema trabalhará com os testes de verificação e integridade, coleta de dados, variação de dados, eventual swap de disco e etc.

4. Considerações finais

Após a descrição da análise, foi possível entender parte do funcionamento do sistema de preservação digital LOCKSS operando em uma rede brasileira com nós geograficamente dispersos.

Foi também possível ver que o processo exige um esforço humano para tratar e padronizar os objetos digitais, assim como implantar, configurar e gerenciar o sistema para que a rede fique operacional.

É importante frisar sobre o cálculo de prospecção para o crescimento da rede, uma vez que as instituições necessitam desta previsão para preparar o ambiente adequadamente.