Modelo de Preservação Hipatia: metodologia de estudo de metadados para extração

Hipatia preservation model: study case of the metadata extraction methodology

Ívina Flores Melo[(1)], Tatiana Canelhas[(2)], Tiago Emmanuel Nunes Braga[(3)]

(1) Instituto Brasileiro de Informação em Ciência e Tecnologia, ivinamelo@ibict.br
(2) Instituto Brasileiro de Informação em Ciência e Tecnologia, tatianacanelhas@ibict.br
(3) Instituto Brasileiro de Informação em Ciência e Tecnologia,tiagobraga@ibict.br

Resumo:
Trata-se de trabalho qualitativo descritivo que tem por objetivo descrever, brevemente, a metodologia de estudo de metadados do Modelo de Preservação Hipatia. O modelo de preservação Hipatia, desenvolvido e pesquisado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia aplica o Modelo OAIS em ambientes digitais institucionais e é uma camada de barramento tecnológico que automatiza o processo de preservação digital. Como objetivos específicos pretende-se relatar o processo de extração de dados executado nos projetos do modelo de preservação Hipatia e apresentar um case (Case Arquivo Nacional). O relato do processo de extração dá-se pela explicação das sete etapas necessárias para a elaboração do dicionário de dados. O relato do estudo de caso demonstra a aplicação das etapas de estudo de metadados. A partir do modelo de referência OAIS, empreende-se esforços na formatação de ambientes sistêmicos e interoperáveis de maneira a viabilizar a preservação digital. Por meio do Hipatia, é possível vislumbrar que o OAIS possa ganhar espaços e amadurecer o cenário brasileiro na temática.

Palavras-chave: Modelo de Preservação Hipatia. Extração de dados. Preservação Digital.

Abstract:
This is a qualitative research that aims to describe the metadata study methodology used by the Hipátia preservation model. The Hipátia preservation model is a research project developed by the Instituto Brasileiro de Informação em Ciência e Tecnologia and is based on the OAIS reference model on digital institutional environments. and developed and researched by the Brazilian Institute of Information in Science and Technology, applies the OAIS Model in institutional digital environments and it is a technological layer that automates the digital preservation process. As specific objectives, we intend to report the data extraction process performed in the projects of the Hipatia preservation model and present a case (National Archive Case). The extraction process is made by the seven steps which are necessary to write down a data dictionary. The case report demonstrated the application of the metadata study stages. Based on the OAIS reference model, efforts are made to format systemic and interoperable environments in order to enable digital preservation. Through Hipatia, it is possible to envision that the OAIS can be used by the Brazilian scenario.

Keywords: Hipatia Preservation Model. Data extraction. Digital Preservation

1 Introdução

Na atualidade, nota-se a crescente e exponencial produção de informação e de desinformação em meios digitais. Neste universo, observa-se que a volumetria de tais informações e, consequentemente, dados produzidos em sistemas da informação demandam tratamento adequado de maneira a torná-los acessíveis ao longo do tempo e, sobretudo, autênticos e confiáveis. Este processo de tratamento, no geral, tem seus pilares no modelo de referência *Open Archival Information System* (OAIS) publicado pela ISO 14.721 (CCSDS, 2012). A exemplo disto, cita-se estratégias internacionais que se apropriam do modelo OAIS tal como o *Digital Curation Life Cycle* do Digital Curation Centre (DCC), situado no Reino Unido, o *Institutional Repository model* estudado pela Universidade de Southampton, também no Reino Unido, e o *Comprehensive Digital Preservation Services* (CDPS) liderado pelo *Massachusetts Institute of Technology* (MIT), dos Estados Unidos da América.

De maneira semelhante, no Brasil tem-se utilizado o mesmo modelo de referência (OAIS) visando a preservação digital. (SARAMAGO, 2004, LIMA et al, 2012) Neste

caminho, as instituições trabalham arduamente na elaboração de estratégias de preservação digital que contemplem os pressupostos do modelo OAIS, o que necessariamente resulta na utilização de *software* de empacotamento, repositórios confiáveis e plataforma de acesso. Este ecossistema tem em seu *core* padrões de empacotamento, tais como o METS, o PREMIS e o EAD, que orientam a normalização dos processos necessários à confiabilidade e autenticidade dos objetos preservados. Porém, esta não é uma tarefa simples. Atender à preservação digital do início ao fim, da gênese ao acesso, cumprindo todos os requisitos técnicos e tecnológicos, demonstra-se uma tarefa de grande empenho.

O modelo Hipatia, proposto pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), foi concebido a partir de uma parceria técnica iniciada em 2018 com o Tribunal de Justiça do Distrito Federal e Territórios (TJDFT) e tinha como objetivo definir uma camada de barramento tecnológico interoperável para garantir a segurança e o acesso aos documentos digitais. Da mesma forma, buscava automatizar o processo de preservação digital. Com o avançar dos estudos, a pesquisa foi ampliada para a proposição de um modelo que visasse a operacionalização e aplicação do Modelo OAIS em ambientes digitais institucionais, proporcionando aplicação técnica e viabilizando o uso de tecnologias na preservação digital. Este modelo foi delineado no Hipatia.

A aplicação do modelo Hipatia é estruturada em cinco fases: preparação arquivística, preparação computacional, extração e empacotamento de objetos digitais, preservação e disseminação. (BRAGA, 2022). Resumidamente, a preparação arquivística diz respeito à análise do sistema produtor e os objetos informacionais produzidos. Esta primeira etapa serve como suporte para a etapa de extração e empacotamento dos objetos digitais, quando é elaborada uma proposta de extração e de organização de objetos, dados e metadados. A preparação computacional baseia-se na análise de duas perspectivas, a primeira, o sistema produtor de objetos informacionais digitais, a segunda, a infraestrutura de rede da instituição. Nesta etapa são estabelecidas as diretrizes para a instalação dos software propostos pelo modelo de preservação, bem como sua configuração. Como resultado desta etapa é proposto um modelo de extração dos dados e metadados e a organização e configuração dos ambientes a fim de se manter a segurança necessária para aplicação do modelo OAIS.

Na etapa de extração de metadados são realizadas as conexões entre o ambiente de produção de objetos digitais e o ambiente de preservação. (BRAGA, 2022). Os objetos e metadados são extraídos do sistema produtor, processados, organizados em um pacote do tipo Bagit e então encaminhados para a etapa de preservação. Na interlocução entre as etapas, o estudo dos metadados do sistema de origem, ou sistema produtor, é colocado como recurso transversal, sendo contemplado pelas preparações arquivísticas e computacionais e caracterizando a extração de dados, bem como subsidiando o envio para preservação.

Sendo assim, este trabalho tem por **objetivo** descrever a metodologia proposta pelo modelo Hipatia para se fazer os estudos de metadados. Como objetivos específicos, pretende-se relatar o processo de extração de dados executado nos projetos do modelo Hipatia e apresentar o caso da pesquisa ocorrida em parceria com o Arquivo Nacional.

Este trabalho caracteriza-se por ter uma abordagem descritiva e qualitativa, bem como traz aspectos de estudo de caso.

3 Metodologia de Estudo

A conjugação das cinco etapas do modelo Hipatia resultam na extração do objeto digital e seus metadados que são enviados para o ambiente de preservação, processados em pacotes com base no Modelo OAIS. Este procedimento objetiva a preservação dos objetos digitais por longos prazos, garantindo-se sua integridade, autenticidade e segurança jurídica. O procedimento de extração, todavia, demanda que os responsáveis envolvidos tenham conhecimento aprofundado dos sistemas

produtores destes objetos digitais. Tal conhecimento perpassa pelo escopo do sistema, suas funcionalidades, a arquitetura do sistema, a arquitetura do banco de dados e os aspectos relacionados ao armazenamento dos objetos e seus formatos resultantes do uso do sistema.

Para se obter esta visão ampliada do sistema é necessário disponibilizar os acessos completos à equipe responsável pela implementação do modelo Hipatia. Não obstante, é essencial que se tenha acesso livre à toda a infraestrutura computacional, como códigos, banco de dados e diretórios de arquivos. Estes acessos são utilizados para se entender a estrutura do sistema e suas funcionalidades, bem como a dinâmica de armazenamento dos dados. É importante observar que neste momento a documentação do sistema pode ser também fonte de consulta, assim como entrevistas com gestores negociais e desenvolvedores.

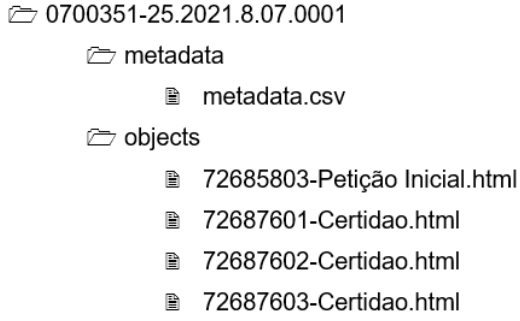
Uma vez realizado o acesso ao sistema produtor, destaca-se o escopo do sistema e inicia-se uma pesquisa em dados, informações e documentos decorrentes do escopo. Procura-se, neste momento, mapear todo o ciclo de vida da produção informacional. Este mapeamento é realizado tanto do ponto de vista da tecnologia da informação quanto do ponto de vista do usuário que busca identificar a proveniência, os contextos e as origens dos metadados descritivos.

A próxima etapa é a de extração de dados, na qual se mapeia a localização dos dados, informações e documentos em servidores locais, também conhecidos como *Local Filesystems*, em *Network Filesystems* (NFS) ou até codificados em banco de dados, do tipo *blob*, por exemplo. Nessa fase, é necessário manifestar o dado, a informação ou o documento tal qual é apresentado no sistema produtor, com todos os seus elementos internos e externos, compostos por formato, conteúdo, metadados descritivos, assinaturas, dentre outros.

Por fim, elabora-se um documento completo, intitulado Dicionário de Dados, para que a equipe de desenvolvedores possa programar no barramento que é utilizado durante a etapa de extração e empacotamento, o BarraPres. O BarraPres realiza a integração entre o sistema produtor e o ambiente de preservação. Ele executa a extração, preparo, organização de pacotes iniciais e transposição de um ambiente para outro, bem como mantém preservadas informações relacionadas ao funcionamento do modelo, tais como *logs* de ações e registros em banco.

Uma vez iniciada a extração, ela é organizada pelo BarraPres em um padrão identificável pelo sistema de preservação, com uso dos diretórios *metadata* e *objects*.

Figura 1: Dados extraídos e sistematizados pelo BarraPres



Fonte: elaboração própria

Após a extração, o Barrapres migra a pasta que foi salva com os dados que serão preservados para o formato *BagIt*[1]. Este pacote gerado é denominado Pacote de Transferência Inicial (PTI) e é apresentado na Figura 2.

[1] *BagIt* é uma padrão convencionado pela *Library of Congress* para empacotar diretórios de arquivos gerando e registrando checksums para cada arquivo armazenado em uma bag, possibilitando a verificação da integridade dos arquivos. Uma bag é o nome do file system directory que contera minimamente data, manifest file com MD5 checksum e um bagit.txt.

Figura 2: Pacote de Transferência Inicial

Fonte: elaboração própria

Após a formatação do PTI, o BarraPres utiliza uma API-Rest para enviar o pacote ao sistema de gerenciamento do empacotamento de repositório Archivematica, de forma sistêmica quando se inicia o empacotamento e a formação do *Submission Information Package* (SIP) preconizado pelo Modelo OAIS. A API utilizada é a */api/v2beta/package*[2], que é utilizada pelo BarraPres por ser a única que possui o parâmetro de endereçamento do sistema de acesso denominado *access_system_id*. Na figura 3, segue exemplo da lista de parâmetros dessa API. O parâmetro *path* é o caminho onde está mapeado o PTI que foi convertido para base64 anteriormente.

Figura 3: Parâmetros da API de transferência

Body raw (json)

json


```
{
  "name": "0700351-25.2021.8.07.0001",
  "path": "L2hvbWUvdWJ1bnR1LzA3MDAzNTEtMjUuMjAyMS44LjA3LjAwMDE=",
  "type": "zipped bag",
  "processing_config": "automated",
  "accession": "",
  "access_system_id": "slug do AtoM",
  "auto_approve": true
}
```


Fonte: elaboração própria

A partir do que foi descrito nesta seção e em síntese, o estudo para extração se apresenta nas seguintes etapas:

1. Acesso ao sistema produtor e sua documentação, se for o caso;
2. Estudo do escopo e da produção informacional;
3. Mapeamento da localização de dados, informações e documentos;
4. Elaboração do dicionário de dados;
5. Configuração do BarraPres;
6. Extração de dados e criação do PTI
7. Envio do PTI para o ambiente de Preservação.

4 Estudo de caso do Arquivo Nacional

Em 2021, o IBICT firmou parceria técnica com Arquivo Nacional (AN), que teve em seu escopo principal a busca pela preservação de processos produzidos no Sistema Eletrônico de Informações (SEI). No âmbito metodológico, o estudo foi feito em um ambiente de homologação criado pelo AN somente para uso do IBICT. Selecionou-se alguns processos como exemplo. Com os números de processos, denominados NUP, em mãos, foi feita a consulta dos metadados do processo e seus objetos, documentos pertencentes ao processo, na interface de usuário do SEI. Nesta primeira consulta, os principais metadados são manifestados nas telas, como dados do processo e toda sua movimentação. O sistema utilizado pelo usuário final foi objeto de primeiro estudo

[2] Disponível em https://www.archivematica.org/en/docs/archivematica-1.13/dev-manual/api/api-reference-archivematica/#package

para conhecimento da criação documental e todos os seus trâmites do ciclo de vida.

Em seguida, mapeou-se os metadados encontrados nas interfaces e buscou-se quais retornavam pelas webservices (WB) do SEI. Para aqueles metadados que não eram recuperáveis por WB, usou-se o acesso direto ao banco de dados, consultando dados como número do processo, interessados, assunto, nível de acesso, tipo de processo do SEI. Para isso foi utilizado o MySQL Workbench 8.0 CE.

Figura 4: Exemplo de consulta ao banco dos interessados do documento

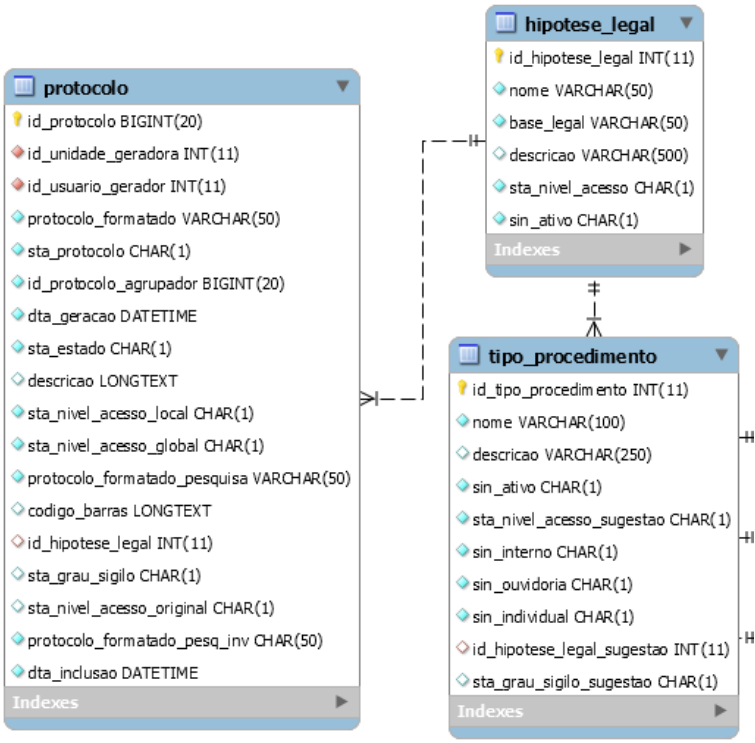
```
SELECT distinct c.sigla,
   c.nome,
   case
      p.sta_participacao
      when 'I' then 'Interessado'
      when 'D' then 'Destinatário'
      when 'R' then 'Remetente'
      when 'A' then 'Acesso Externo'
      else p.sta_participacao
   end as "destinatario"
FROM participante p
   INNER JOIN contato c on p.id_contato = c.id_contato
WHERE sta_participacao = 'I'
   and ID_PROTOCOLO in (
      SELECT p2.id_protocolo
      FROM protocolo p
         inner join rel_protocolo_protocolo rpp on p.id_protocolo =
               rpp.id_protocolo_1
         inner join protocolo p2 on rpp.id_protocolo_2 = p2.id_protoc
      WHERE p.id_protocolo = id_protocolo //recuperado em consulta anterior
   );
```

Fonte: elaboração própria

Além dos metadados, os documentos também devem ser recuperados tal como são manifestados em tela para o usuário, em sua formatação e visualização dos principais elementos como dados da assinatura digital. Durante o estudo, percebeu-se que os documentos produzidos pelo sistema SEI são códigos *.html* salvos em diferentes campos e tabelas do banco de dados, mas todo seu conteúdo pode ser recuperado em um só campo retornado pela WB "ConsultaDocumento.LinkAcesso"[3]. Já os arquivos que são anexados aos processos ficam salvos em um *filesystem* local, mas sua recuperação acontece da mesma forma que os documentos citados anteriormente, ou seja, utilizando-se um WB fornecido pelo SEI.

Além das pesquisas diretas nas interfaces do sistema, no banco de dados e nas WS, outro método de acesso do sistema e mapeamento dos metadados deu-se apoiado por meio do Modelo de Entidade Relacionamento (MER) do SEI, possibilitando mapear outras informações tal como é apresentada na figura 3, a seguir:

Figura 3- Exemplo de MER localizada



Fonte: elaboração própria

Com os metadados de gestão mapeados, comparou-se o que pode ser recuperado do SEI com os metadados citados pelo e-Arq Brasil, modelo de requisitos para sistemas de gestão arquivística de documentos, publicado pelo Conselho Nacional de Arquivos (Conarq) para órgão do Poder Executivo Federal. Coube ao AN escolher quais destes metadados seriam preservados, mesmo que eles não fossem citados pelo modelo. Aproveitou-se essa seleção também para

[3] https://www.gov.br/economia/pt-br/assuntos/processo-eletronico-nacional/arquivos/documentacao-do-sei/sei-webservices-v3-1.pdf

compor alguns metadados descritivos no padrão de interoperabilidade *Dublin Core* (DC), utilizado entre o *Archivematica*[4] e o *AtoM*[5]. O formato usado para salvar os metadados no projeto do Arquivo Nacional foi o *Comma-separetad Value* (CSV), respeitando as regras de disponibilização dos dados conforme documentação do *Archivematica*. O primeiro campo deve, obrigatoriamente, ser nomeado como *filesystem*. Os seguintes foram os campos DC, seguidos dos metadados de gestão. O documento CSV foi salvo em codificação UTF-8 e teve seus valores separados por vírgulas.

Com o estudo finalizado, consolidou-se toda a informação em um documento voltado para a equipe de desenvolvimento do projeto, o Dicionário de Dados, que descrevia quais os metadados e os documentos seriam preservados, como extraí-los do SEI e como seriam disponibilizados em diretórios específicos para serem salvos no formato *BagIt*.

4 Conclusão ou Considerações Finais

A partir do modelo de referência OAIS, empreendeu-se esforços na formatação de ambientes sistêmicos e interoperáveis de maneira a viabilizar a preservação digital. Um processo complexo e cujas demandas exigiram estudos que viabilizassem a aplicação do modelo de referência OAIS.

O modelo de preservação Hipatia, proposto pelo Instituto Brasileiro de Informação em Ciência e Tecnologia, aplica o Modelo OAIS em ambientes digitais institucionais e busca mediar os processos de preservação digital, tendo suas bases nas melhores práticas em cinco etapas.

Pode-se observar que o conhecimento do sistema produtor é essencial ao uso de tecnologias que viabilizem a preservação digital Foi também visto que ela apenas se consolidará com a estruturação de um ecossistema integrado mediado por padrões e normas. Por meio do Hipatia, foi possível vislumbrar que o Modelo de referência OAIS pode se tornar realidade corrente nos cenários informacionais do Brasil e se apresenta como uma solução para atender a demanda latente de implementação de Repositórios Arquivísticos Digitais Confiáveis.

Referências

BRAGA, Tiago E N. O modelo Hipatia: a proposta do IBICT para a preservação digital arquivística. **In: Hipatia: modelo de preservação para repositórios arquivísticos digitais confiáveis.** Brasília: Ibict, 2022. Disponível em: <http://labcotec.ibict.br/omp/index.php/edcotec/catalog/book/livroHipatia>. Acesso em: 30 ago 2022.

CCSDS – Consultative Committee for Space Data Systems. **Reference Model for an Open archive Information System (OAIS).** Washington: CCSDS Secretariat, June 2012. Disponível em: . Acesso em: 09 set 2022.

CONSELHO NACIONAL DE ARQUIVOS. **Modelo de requisitos para sistemas informatizados de gestão arquivística de documentos: e-ARQ Brasil.** Versão 2. Rio de Janeiro: Arquivo Nacional, 2022. Disponível em: <https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/EARQV203MAI2022.pdf>. Acesso em 09 set 2022.

LIMA, Rodrigues de Souza, A. H., OLIVEIRA, A. F., D'AVILA, R. T., CHAVES, E. P. da S. S. (2014). O modelo de referência OAIS e a preservação digital distribuída. **Ciência Da Informação, 41**(1). https://doi.org/10.18225/ci.inf.v41i1.1352 Acesso em: 07 set 2022

[4] Archivematica é um software de empacotamento e preservação segundo o Modelo OAIS e gerenciamento de repositório confiável. É desenvolvido pela Artefectual.

[5] AToM (acesso to memory) é uma plataforma de acesso e difusão, aderente ao Modelo OAIS, desenvolvido pela Artefectual e cujo escopo debruça-se na disponibilização de representantes/derivadas de documentos preservados.

SARAMAGO, Maria Lurdes. Metadados para preservação digital e aplicação do modelo OAIS. In: **Actas do congresso nacional de bibliotecários, arquivistas e documentalistas.** 2004.