11ª conferência luso-brasileira
CIÊNCIA ABERTA
braga-portugal
6 a 8 outubro 2020

Análise de níveis de preservação digital para repositórios institucionais

Andréa Gonçalves, ICICT/Fiocruz
Claudete Queiroz, ICICT/Fiocruz
Luciana Danielli, ICICT/Fiocruz

ICICT Instituto de Comunicação e Informação Científica e Tecnológica em Saúde
Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
SUS
MINISTÉRIO DA SAÚDE
GOVERNO FEDERAL

Alguns desafios para RIs no Brasil

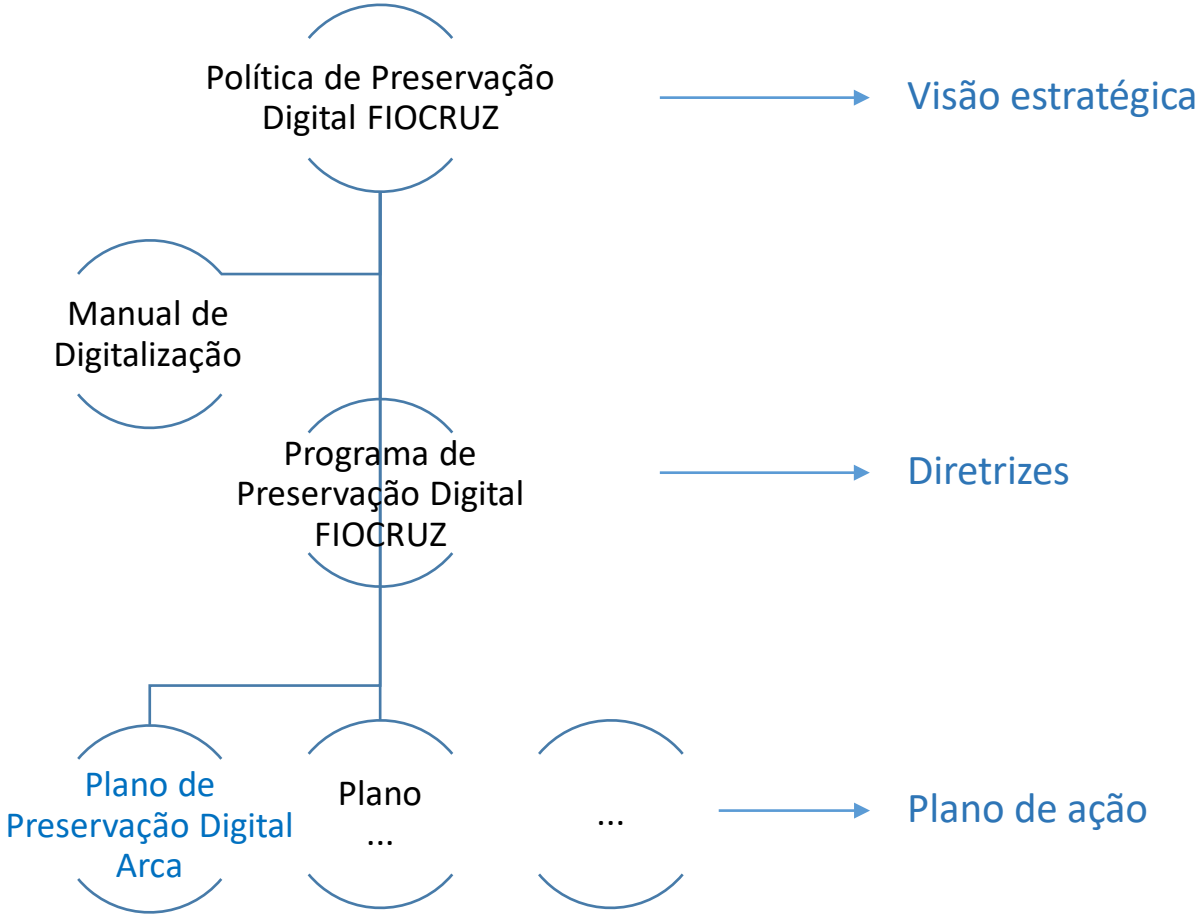
- Infraestrutura tecnológica e capacitação
- Documentação de procedimentos e práticas
- Autoavaliação / auditoria de confiabilidade
- Políticas e planos de preservação digital

Repositório Institucional Arca

- Preservação da memória técnico-científica da Fundação Oswaldo Cruz (FIOCRUZ), desde 2011
- Possui cerca de 40.000 registros

- Recomendações de confiabilidade: instituição de estratégias documentadas de preservação, implementada e continuada.
- Plano de Preservação Digital do Arca

Contexto institucional



Programa de Preservação Digital da Fiocruz

- Requisitos mínimos para aprovação do Plano de PD:
  - Armazenamento redundante
  - Verificação periódica de vírus
  - Inventário dos arquivos digitais

- Incluir em revisão futura auditoria e certificação do modelo OAIS, bem como a avaliação dos **níveis de preservação digital**, com base em padrões e recomendações internacionais

Modelos de níveis de preservação digital

1. National Digital Stewardship Alliance (NDSA)
2. Universidade de Alberta (UA)
3. Universidade de British Columbia (UBC)
4. Sustainable Heritage Network (SHN).

NDSA

Functional Area	Level			
	Level 1 (Know your content)	**Level 2 (Protect your content)**	**Level 3 (Monitor your content)**	**Level 4 (Sustain your content)**
Storage	Have two complete copies in separate locations Document all storage media where content is stored Put content into stable storage	Have three complete copies with at least one copy in a separate geographic location Document storage and storage media indicating the resources and dependencies they require to function	Have at least one copy in a geographic location with a different disaster threat than the other copies Have at least one copy on a different storage media type Track the obsolescence of storage and media	Have at least three copies in geographic locations, each with a different disaster threat Maximize storage diversification to avoid single points of failure Have a plan and execute actions to address obsolescence of storage hardware, software, and media
Integrity	Verify integrity information if it has been provided with the content Generate integrity information if not provided with the content Virus check all content; isolate content for quarantine as needed	Verify integrity information when moving or copying content Use write-blockers when working with original media Back up integrity information and store copy in a separate location from the content	Verify integrity information of content at fixed intervals Document integrity information verification processes and outcomes Perform audit of integrity information on demand	Verify integrity information in response to specific events or activities Replace or repair corrupted content as necessary
Control	Determine the human and software agents that should be authorized to read, write, move, and delete content	Document the human and software agents authorized to read, write, move, and delete content and apply these	Maintain logs and identify the human and software agents that performed actions on content	Perform periodic review of actions/access logs
Metadata	Create inventory of content, also documenting current storage locations Backup inventory and store at least one copy separately from content	Store enough metadata to know what the content is (this might include some combination of administrative, technical, descriptive, preservation, and structural)	Determine what metadata standards to apply Find and fill gaps in your metadata to meet those standards	Record preservation actions associated with content and when those actions occur Implement metadata standards chosen
Content	Document file formats and other essential content characteristics including how and when these were identified	Verify file formats and other essential content characteristics Build relationships with content creators to encourage sustainable file choices	Monitor for obsolescence, and changes in technologies on which content is dependent	Perform migrations, normalizations, emulation, and similar activities that ensure content can be accessed

TAP - Universidade de Alberta (UA)

Table 1: Types of resources

Type of Resource	Score
Collection of Strength	5
Local Born Digital Resources	4
Purchased / Digitized Resources	3
Licensed Resources	2
External Resources	1

Table 2: Scoring Levels for Archival Responsibility

Archival Responsibility	Score
Sole Responsibility	2
Shared Responsibility	1
Third-party Responsibility	0

Table 3: Example of Projected Preservability

File Format	Adoption	Openness	Transparency	Stability	Interoperability	%	PP	Score
xml	2	2	2	2	2	100%	High	3
pdfa	2	2	2	2	2	100%	High	3
rtf	1	0	1	2	2	60%	Medium	2
bmp	1	0	2	0	0	30%	Low	1

Table 4: Example of a Value Matrix

Type of Resource	Archival Responsibility	Projected Preservability	%	Level
5	2	3	100%	Gold
4	2	2	80%	Silver
2	1	1	40%	Bronze

Universidade de British Columbia (UBC)

	Level 1: Basic Preservation	Level 2: Bit-level Plus Preservation[1]	Level 3: Full Preservation
Type of content	❑ external digitization requests ❑ legacy digitized content ❑ selected/licensed research data sets ❑ in copyright material ❑ file format conversion projects ❑ licenced data sets	❑ other locally digitized resources (e.g., retrospectively scanned newspapers) ❑ low quality files ❑ material of lower projected preservability	❑ flagship digitization projects representing collections of local strength ❑ locally created born digital collections ❑ externally created resources for which we have stewardship responsibilities i.e. Chinese Canadian Stories Community Collections ❑ COPPUL PLN content (200 GB) ❑ CGI PLN content (consortial) ❑ select research data sets (DataVerse)
Storage and Geographic Location	❑ 2 complete copies ❑ transfer from heterogeneous media to storage system	❑ 3 complete copies ❑ 1 copy in different geographic location ❑ document storage system	❑ crash consistent snapshot is taken every morning at 3am and vaulted over to a remote location at midnight the same day

Sustainable Heritage Network (SHN)

Levels of Digital Preservation Preparedness

This resource is a guide to understanding digital preservation, with a focus on three different levels of preparedness. This resource will help you to start minimal digital preservation steps at your institution, and show how to build upon digital preservation incrementally.

For more information about digital preservation procedures, tools, and policies, view related items connected to this resource on the Sustainable Heritage Network in the "Digital Preservation" category.

- The Three Essentials of Digital Preservation
- Activities to Include in a Digital Preservation Plan
- Digital Preservation Glossary
- Developing a Digital Preservation Policy

Introduction

Digital preservation is a series of activities, plans, and policies combined to ensure continued preservation and access of digital materials. Organization and management of digital files is just as important as any other format type. In fact, digital files can be even more vulnerable than a piece of paper or photograph.

Modelo	Níveis	Categorias
National Digital Stewardship Alliance (NDSA)	Nível 1: Proteja seus dados Nível 2: Conheça seus dados Nível 3: Monitore seus dados Nível 4: Repare seus dados	• Armazenamento • Integridade • Controle • Metadados • Conteúdo
Modelo TAP	Nível Bronze Nível Prata Nível Ouro	• Tipo de recurso (T) • Responsabilidade do Arquivo (A) • Preservabilidade do formato (P)
Biblioteca da Universidade de British Columbia	Nível 1: Preservação básica (bit-level) Nível 2: Preservação intermediária (bit-level plus) Nível 3: Preservação completa	• Tipo de conteúdo • Armazenamento e localização geográfica • Fixidez e integridade • Segurança da informação • Metadados • Formatos de arquivo
Sustainable Heritage Network	Nível Mínimo Nível Intermediário Nível Avançado	• Armazenamento de arquivos • Integridade de arquivos • Acesso a arquivos

Fonte: Autoria própria, 2020

Recomendações

- Definição das categorias/áreas principais (Armazenamento, Integridade, Conteúdo, Acesso, Formato, Metadados) e níveis de operação (Básico, Intermediário, Avançado)
- Detalhamento dos critérios e ações mínimas exigidas em cada nível
- Relacionar resultados às recomendações de confiabilidade para repositórios institucionais (ISO 16363): Infraestrutura organizacional, Gestão dos objetos digitais, Infraestrutura tecnológica e de segurança

Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

www.facebook.com/fiocruz.icict

twitter.com/@Icict_fiocruz

www.youtube.com/videosaudefio

www.icict.fiocruz.br