Projeto e Preservação de Repositórios Digitais

Marcos Sunye

Universidade Federal do Paraná

C3SL

Centro de Computação Científica e Software Livre

C3SL

- Centro de Computação Cientifica e Softwate Livre
- Maior espelho da America Latina para projetos de Software Livre
- Maior provedor de conteúdo da rede RNP
- Desenvolvedor do Linux Educacional (200.000 instalações), Portal Pro Info Data, Portal Cidades Digitais, Rede Social Participatório

Preservação Digital

- C3SL e a Preservação Digital
  - Biblioteca Digital da UFPR criada em 2004
    - 3000 videos (TV UFPR)
    - 13000 teses/dissertacoes
    - 38 revistas científicas eletrônicas
    - Comissão da verdade Paraná
    - 3a Biblioteca Digital do Brasil (Webometrics)
  - Acervo Digital do Paraná
    - Jornais e Livros raros da Biblioteca Pública do Paraná

Preservação Digital

- Motivação
  - Bibliotecas Digitais
  - Memória Digital
    - Jornais Eletrônicos
    - Blogs
    - Páginas Institucionais

Preservação Digital

- DESAFIOS
- Disponibilidade
  - Conteúdo digital acessível (em disco)
  - Controle de permissão Acesso (senha)
  - Formato compatível (evolução de versões e distribuições)
- Confiabilidade
  - Conservar o conteúdo intacto através do tempo
  - Auditoria Constante (Arquivos corrompidos)

Preservação Digital

- Equipe que combina Ciência da Informação e Computação

  Problemas relacionados à Ciencia da Informação

  - Qualidade dos metadados (catalogação)
  - Planejamanento do Acervo
  - Priorização
  - Direitos autorais (Livros, Imagens, Videos)

Preservação Digital

- Problemas Relacionados à Computação
  - A quantidade de informação na Internet tem um crescimento acelerado (80% do total do conteúdo foi gerado nos ultimos 2 anos)
  - Ao escalar a quantidade de informação (vários TeraBytes) são necessarios servidores não convencionais (a informação já não cabe em apenas 1 disco)
  - Raids, Array de Discos, Particionamento etc
  - Fonte Redundante, No-Break, Gerador
  - Discos caros

Preservação Digital

- Problemas relacionados à Computação
  - Ao escalar a quantidade de informação (vários TeraBytes) são necessarios recursos humanos especializados
  - Administrador de Sistema
  - Administrador de Banco de Dados

Preservação Digital

- Problemas relacionados à Computação
  - Ao escalar a quantidade de informação (vários TeraBytes) são necessarias estratégias específicas e **recursos permanentes**
  - Escalabilidade dos equipamentos deve acompanhar o crescimento da informação
  - As rotinas de backup podem durar varias horas e devem ser feitas de maneira incremental

Preservação Digital

- Custos associados
  - Hardware extremamente caro
    - ”Storage” de 10Tb = 200.000 uma servidora de igual capacidade custa 10x menos!
  - Ambiente Computacional Adaptado
    - Rede Eletrica e Lógica
  - Recursos Humanos raros (principalmente em Instituições Públicas) e caros

Preservação Digital

- Alternativas para minimizar o problema:
  - Automatizar o processo de backup
  - Terceirização (Data centers)
  - Multiplicar o número de cópias
  - Redes cooperativas

Preservação Digital

- Multiplicar o número de cópias
  - Fazer tantas cópias quantas forem necessárias para garantir a confibilidade
  - Vantagens:
    - Procedimentos mais simples
    - Hardware mais simples
    - Disponibilidade
  - Problemas
    - Segurança

Preservação Digital

- Multiplicar o número de cópias
  - Criar redes cooperativas
  - Projeto Lockss
    - Universidade de Stanford (Décimo ano)
    - Uso de redes Peer to Peer
    - Auditoria e reparo contínuos
    - Máquina dedicada
    - Código Aberto

Preservação Digital

- Multiplicar o número de cópias
  - Redes P2P
    - Vídeos, Música, Jogos
    - Azureus, Emule, Torrent etc..
    - Filosofia de multiplicar as cópias para aumentar a disponibilidade, desempenho de download etc
    - Arquitetura já consolidade
      - DHT (Distributed Hash Table)
      - put/get

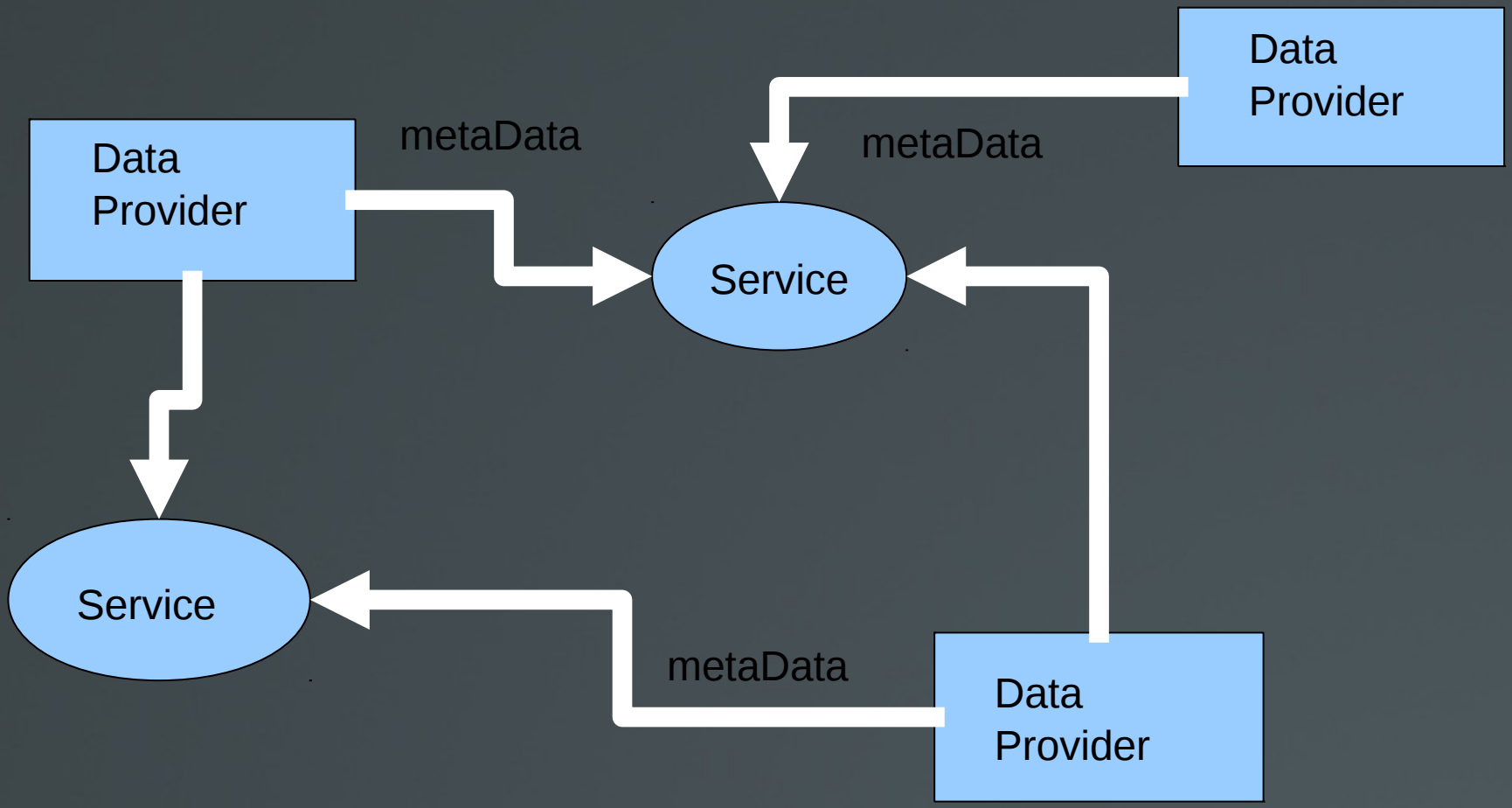
Preservação Digital

- Multiplicar o número de cópias
  - Data Trading
  - Troca de conteúdos entre os Repositórios
  - Autonomia
  - Confiabilidade
  - Disponibilidade

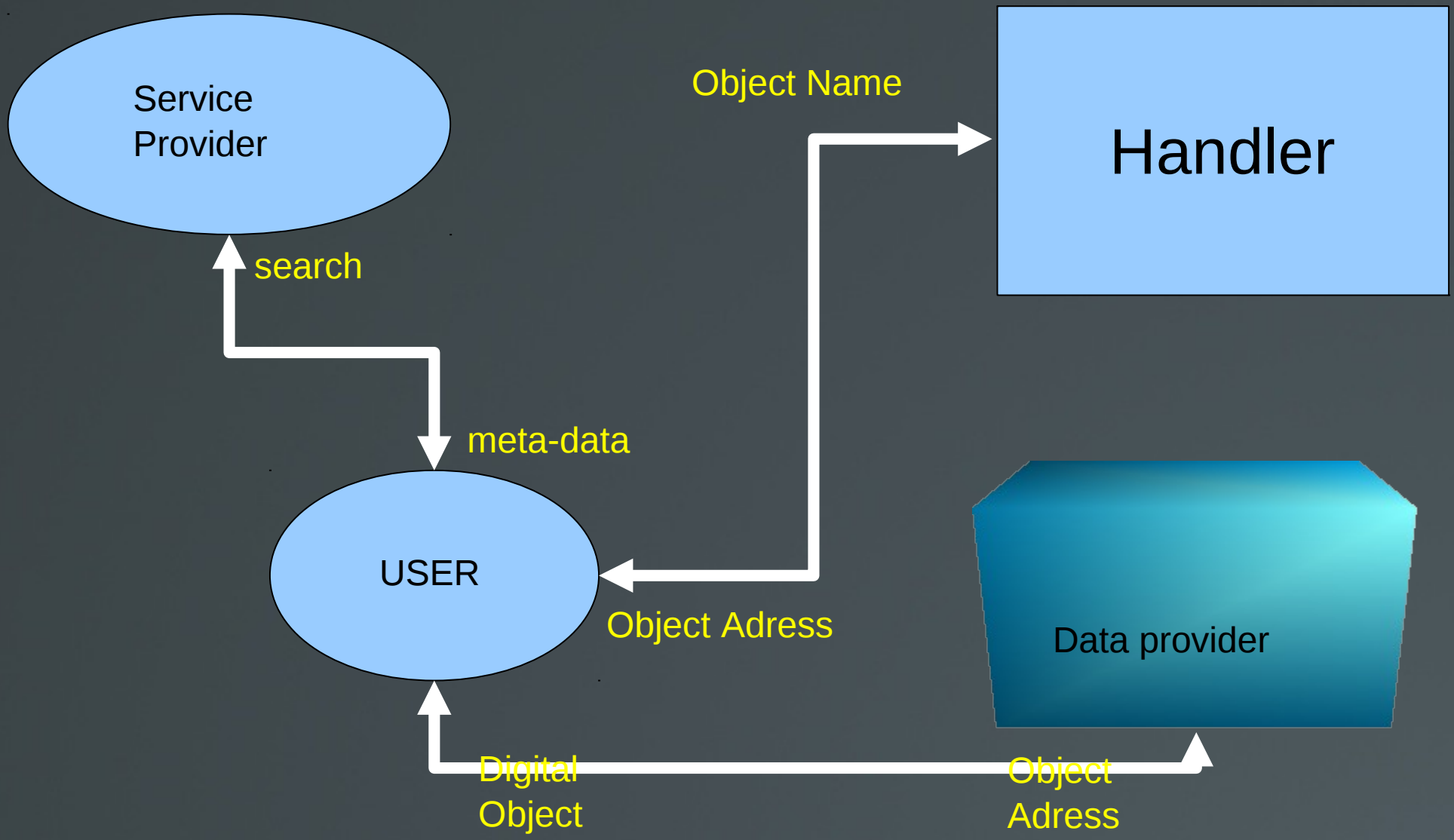
Preservação Digital

- Redes Cooperativas
  - Open Archives Initiative
    - Copias dos metadados
    - Redes consolidadas
    - Cooperação e compatibilidade entre as Bibliotecas Digitais

Preservação Digital



Preservação Digital



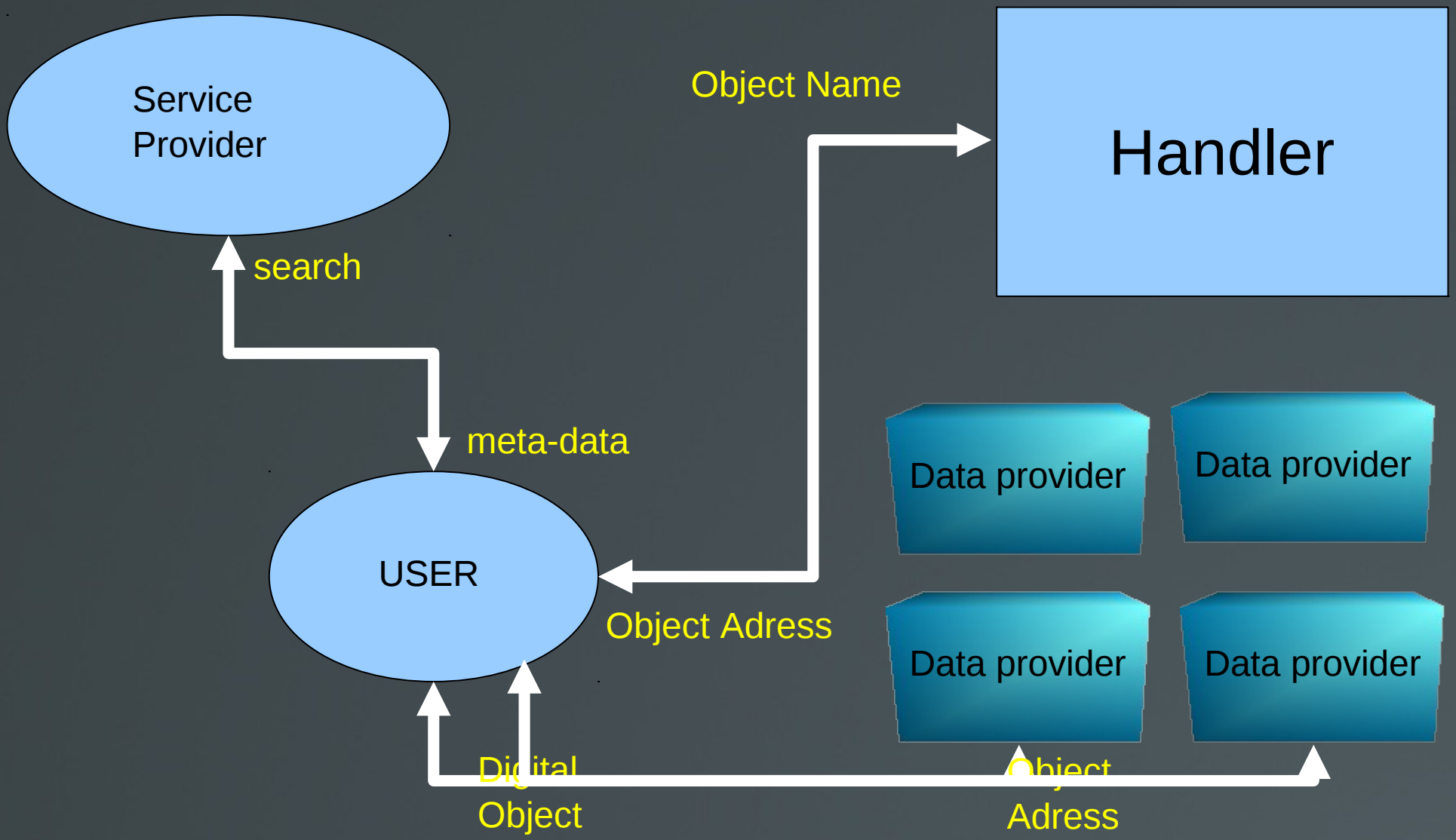
Preservação Digital

- <header> <identifier>oai:ufpr.br:226143</identifier> <datestamp> 2006-11-30 </datestamp> </header>
- <metadata>
- <dc:type>text</dc:type> <dc:language>po</dc:language> <dc:creator>Cancelli, Diana Maria</dc:creator>
- <dc:identifier>http://hdl.handle.net/1884/6799</dc:identifier>
- <dc:title>Um modelo para a evolucao termica de lagos profundos / </dc:title>
- <dc:publisher></dc:publisher> <dc:date>2006.</dc:date> <dc:description>Orientador: Nelson Luis da Costa Dias</dc:description>
- <dc:description>Dissertacao (mestrado) - Universidade Federal do Parana, Setor de Ciencias Exatas e Setor de Tecnologia, Programa de Pos-Graduacao em Metodos Numericos em Engenharia. Defesa: Curitiba, 2006</dc:description> <dc:contributor>Dias, Nelson Luis da Costa </dc:contributor>
- <dc:contributor>Universidade Federal do Parana. Setor de Tecnologia.Setor de Ciencias Exatas.Programa de Pos-Graduacao em Metodos Numericos em Engenharia.</dc:contributor>
- </oai_dc:dc>
- </metadata>

Preservação Digital

- Handle System
  - Permite a independencia entre a maquina/dominio do repositório digital e o endereço do objeto digital
  - Parte do mecanismo que implenta o DOI (Digital Object Identifier).
  - Serviço sem fins lucrativos (cerca de US$ 150/ano)
  - Necessário antes da implantação
  - Pode ser um fator de atraso na implantação

Preservação Digital



Dspace

- Mais de 1000 respositórios institucionais cadastrados (maior quantidade de Data Providers)
- Versao 3.2 disponivel no Source Forge
- Facil instalação (desde que se tenha conhecimento em servidores WEB e SGDB Relacional)
- Apache Tomcat, Postgres ou Oracle

Dspace

- Representação hierárquica
  - Comunidades e coleções
  - Problemas com livros e revistas
- Busca Unificada
- Integração com outros sistemas via OAI/PMH
- Integração com o Handle System

Dspace

- Distribuição estável
- Comunidade muito ativa
- Localização e Internacionalização
- Integração com outras interfaces
  - Biblioteca USP/Brasiliana (Corisco)
  - Projeto 100 anos (UFPR)
  - Comissao da Verdade

Dspace

- O dspace é uma pltaforma de preservação de documentos digitais.
- A definição de interfaces pode ser um projeto a parte.

Preservação Digital

- Conclusão
  - Preservação Digital custa caro
  - Preservação Digital requer infraestrutura não convencional
  - Muitas cópias é uma boa idéia desde que não existam preocupações com a segurança (autoria, modificação de conteúdo não autorizado etc)
  - Dspace é um consenso (externalidade de rede)

Preservação Digital

- Referencias

- 20th International Conference on Advanced Information Networking and Applications - Volume 1 (AINA'06) Freelib: Peer-to-peer-based Digital Libraries
- Defending a P2P Digital Preservation System, Bryan Parno, IEEE Transactions on Dependable and Secure Computing, Volume 1 Issue 14, pgs 208-222. december 2004
- ``A Fresh Look at the Reliability of Long-term Digital Storage'' Mary Baker, Mehul Shah, David S. H. Rosenthal, Mema Roussopoulos, Petros Maniatis, TJ Giuli, Prashanth Bungale, , *Proceedings of EuroSys*, April, 2006.
- Peer-to-Peer Data preservation through Storage Auctions, Brian F Cooper, Hector Garcia-Molina IEEE Transactions on Parallel and Distributed Systems, Vo. 16, NO. 3, March 2005
- www.lockss.org / www.digitalpreservation.gov/