Bibliotecas Digitais e os desafios do Arquivamento Digital, uma visão da ciencia da computação

- Marcos Sunye, Universidade Federal do Paraná
- C3SL (Centro de Computação Científica e Software Livre)

Arquivamento Digital

- C3SL e o Arquivamento Digital
  - Biblioteca Digital da UFPR
    - 3000 videos
    - 3700 teses/dissertacoes
    - 32 revistas científicas eletrônicas
  - Acervo Digital do Paraná
    - Jornais e Livros raros da Biblioteca Pública do Paraná
  - 3000 usuários em 600 terminais
    - WebPage, Email, Area de Trabalho

Arquivamento Digital

- Os Sistemas de Preservação Digital surgiram para atender a demanda na acessibilidade e segurança de documentos digitais no longo prazo (período superior a um século);
- Aquisição, Indexação, Distribuição
- O Arquivamento Digital é o ramo responsável pela integridade dos dados armazenados.

Arquivamento Digital

- Arquivamento Digital deve obter um melhor desempenho no acesso aos documentos, em relação ao arquivamento tradicional;

- Deve garantir o arquivamento a longo prazo utilizando dispositivos com vida útil, atualmente, em torno de 5 anos.

Caracteristicas do Arquivamento Digital

- Definida pela ISO 14721:2003, pela Open Archival Information System;
- Disponibilidade, durabilidade e confiabilidade dos dados
- Disponibilidade
  - Conteúdo digital acessível (em disco)
  - Controle de permissão Acesso (senha)
  - Formato compatível (evolução de versões e distribuições)

Características do Arquivamento Digital

- Confiabilidade
  - Conservar o conteúdo intacto através do tempo
  - Auditoria Constante (Arquivos corrompidos)

- Indolência: não há necessidade de rapidez nas mudanças do sistema.

Ameaças ao Arquivamento Digital

- Falhas na Mídia;
- Falhas de Hardware;
- Falhas de Software;
- Erros de Comunicação;
- Falhas em Serviços de Rede:
- DNS;
- URL Persistentes.

Ameaças ao Arquivamento Digital

- Obsolescência de Mídia e Hardware;
- Obsolescência de Software;
- Erros do Operador;
- Desastres Naturais;
- Ataques Externos;
- Ataques Internos;
- Falhas Econômicas;
- Falhas Organizacionais.

Arquivamento Digital

- Problemas Relacionados à Computação
  - A quantidade de informação na Internet tem um crescimento acelerado (> 20 hexabytes)
  - 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 SCSI

Arquivamento 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

Arquivamento 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

Arquivamento 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 e caros

Arquivamento Digital

- Alternativas para minimizar o problema:
  - Terceirização (Data centers)
  - Automatizar o processo de backup multiplicando copias
  - Multiplicar o número de cópias em redes cooperativas
  - Uso de redes Peer to Peer (P2P)

Arquivamento Digital

- Replicação
  - prós:
    - Aumenta a confiabilidade, disponibilidade, tolerância à falhas e o desempenho do sistema;
    - Hardware mais simples:
  - Contras:
    - Aumenta o uso dos dispositivos de armazenamento e dificulta a segurança e atualização dos dados.

Arquivamento Digital

- Paradigma Peer-to-Peer (Distribuído)
  - Os peers são máquinas que atuam simultaneamente como cliente e servidor;
  - Eles colaboram entre si, utilizando recursos ociosos de computação e armazenamento;
  - Já usado em Vídeos, Música, Jogos (Azureus, Emule, Torrent etc..)

Arquivamento Digital

- P2P
  - Filosofia de multiplicar as cópias para aumentar a disponibilidade, desempenho de download etc
  - Arquitetura já consolidada DHT (Distributed Hash Table) put/get
  - escalabilidade: a capacidade do sistema aumenta com o número de peers participantes;
  - auto-organização: o sistema se organiza sem intervenção e a entrada e saída de peers não altera seu comportamento;
  - tolerância à falhas: não existem pontos únicos de falhas.

Arquivamento Digital

Projeto Lockss, Poucas cópias em redes cooperativas

- Universidade de Stanford (Décimo ano)
- Uso de redes Peer to Peer
- Máquina dedicada
- Código Aberto
- Auditoria reparo contínuos
  - Descobre as falhas latentes; Dois tipos: periódica e oportunística.

Arquivamento Digital

- Farsite, rede não confiáveis
  - Sistema de arquivos seguro em ambiente não confiável;
  - Projetado para executar em grandes corporações ou universidades, numa rede de até 105 peers (topologia da rede pode ser ignorada);
  - Segue o princípio de que boa parte da capacidade dos discos permanece ociosa;
  - Utiliza capacidade de processamento para proteger o conteúdo com criptografia;

Arquivamento Digital

- Glacier, alto indice de replicação
  - Sistema de armazenamento distribuído projetado para a sobrevivência de falhas correlacionadas em larga escala;
  - Utiliza redundância massiva;
  - Em um teste, alcancou durabilidade de 99,99999% apesar de 60% de falhas correlatas entre os peers.

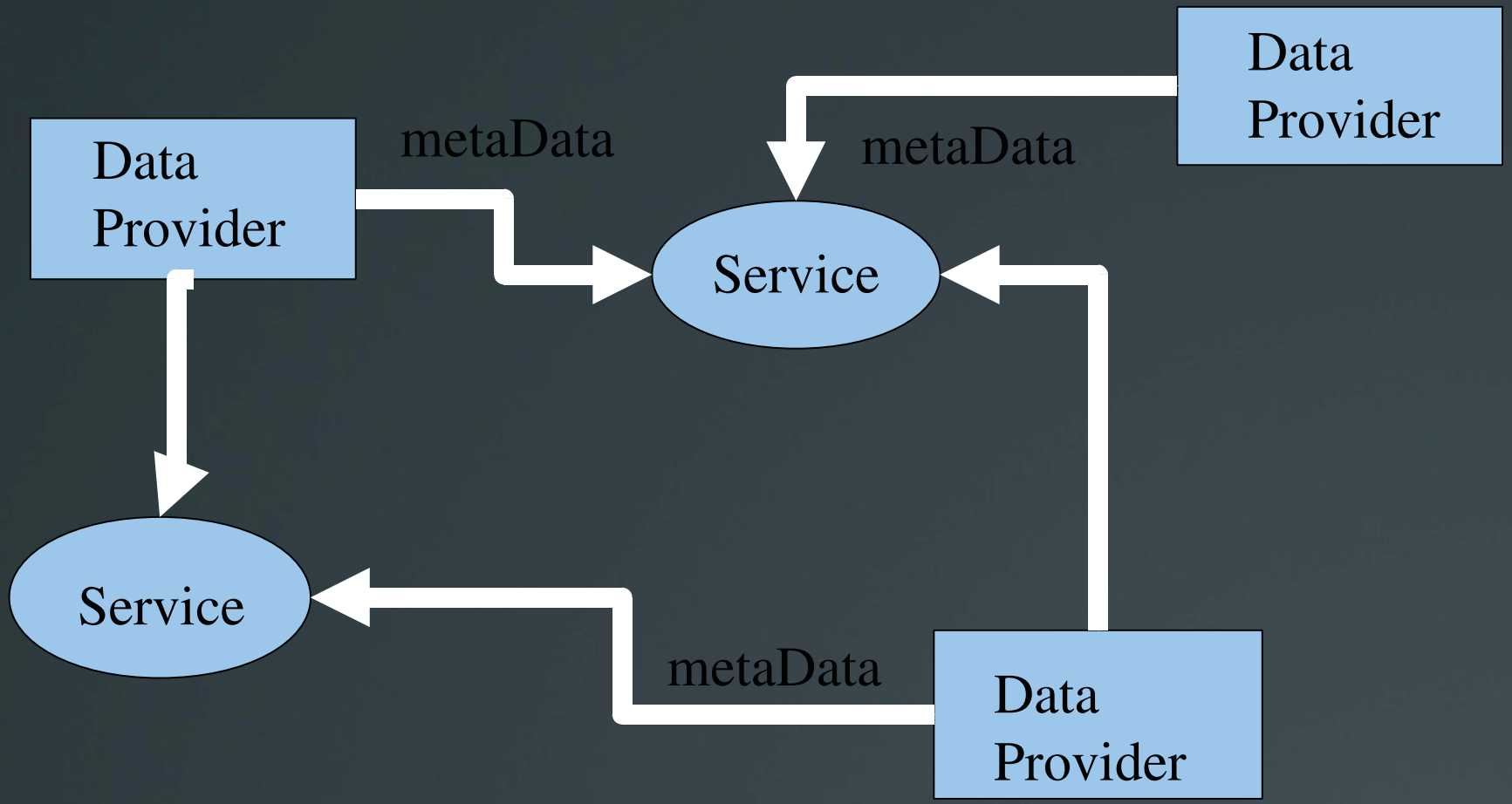
Arquivamento Digital

- Open Archives Initiative
  - Fundada em 2001 derivada da ArXiv (física)
  - Separação entre Metadados e Objetos
  - Distribuição e Cópias dos metadados
  - Redes de cooperação consolidadas e comunidade ativa
  - Cooperação e compatibilidade entre as Bibliotecas Digitais
  - Identificadores de Objeto Digital permanentes (DOI)

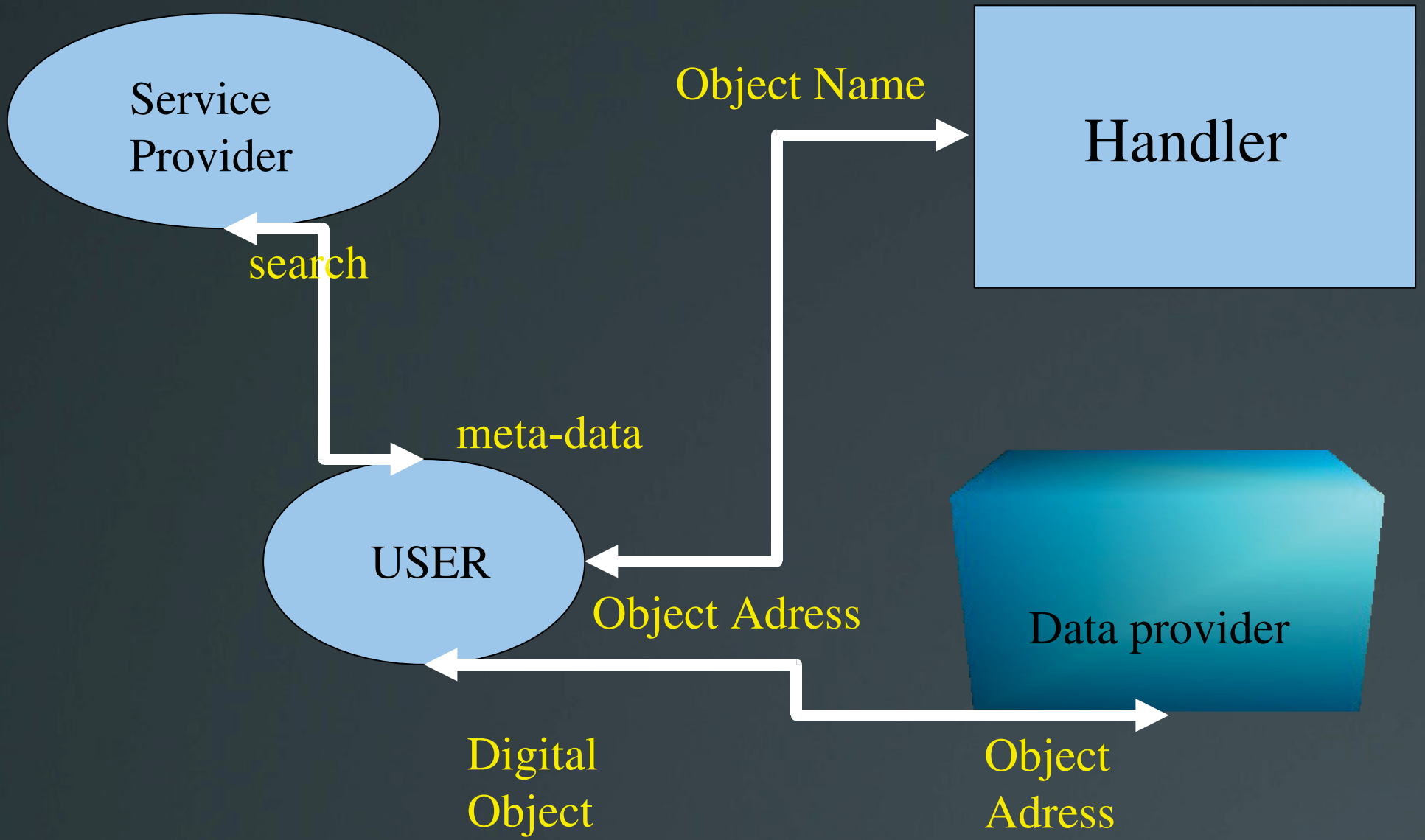
Exemplo ArXiv

- “Thurston's geometrization conjecture, including the Poincaré conjecture”publicado por Grigori Perelman em November 2002.
- Perelman nem se preocupou com a publicação em revistas científicas tradicionais:
  - “ Se alguém esta interessado na minha maneira de resolver o problema esta tudo ali no Arxiv”

Arquivamento Digital



Arquivamento Digital



- Campo DC Valor Idioma
- dc.contributor.author Watter, Leslie Harlley pt_BR
- dc.date.accessioned 2006-08-09T15:21:24Z -
- dc.date.available 2006-08-09T15:21:24Z -
- dc.date.created 2006-08-08 pt_BR
- dc.date.issued 2006 pt_BR
- dc.identifier.uri http://hdl.handle.net/1884/4333 -
- dc.description.abstract Orientador: Roberto A.Hexsel pt_BR
- dc.description.abstract Inclui apendice pt_BR
- dc.description.abstract Dissertacao (mestrado) - Universidade Federal do Parana, Setor de Ciencias Exatas, Programa de Pos-Graduacao em Informatica. Defesa: Curitiba, 2006 pt_BR
- dc.description.abstract Inclui bibliografia pt_BRo Kernel do Linux / pt_BR
- Aparece nas Coleções: Teses & Dissertações

OAI – Dublin Core

- dc.description.abstract Resumo: Para suprir a necessidade de poder computacional de aplicacoes cientificas utilizam-se aglomeradosde computadores. A comunicacao das aplicacoes nesses aglomerados e feita usandobibliotecas de troca de mensagens que utilizam normalmente os protocolos TCP/IP como meiode transporte. Restringindo a rede do aglomerado de computadores a uma rede local e possivel substituir os protocolos TCP/IP pelo protocolo LLC, com ganho de desempenho.Este trabalho apresenta uma avaliacao de desempenho de uma modificacao da biblioteca OPENMPI para trabalhar com o LLC, comparando-a com a implementacao TCP/IP. Para a avaliacao de desempenho foram utilizados os aplicativos NETPIPE, MPPTEST, a Transformada Rapida de Fourier e a ordenacao Radix. Os resultados obtidos para o NETPIPE mostram que o LLC tem um desempenho de 16 a 21% superior ao TCP/IP; para o MPPTEST esse resultado varia na faixa de 3 a 12%, sendo que os maiores ganhos ocorrem para mensagens pequenas. Os resultados da Transformada Rapida de Fourier (FFT) mostram que o LLC e de 2.8 a 6.8% mais rapido que o TCP/IP variandoo numero de processadores de 2 a 16. O resultado da ordenacao Radix nao aponta ganhoreal para o LLC porque esse programa nao gera demanda significativa sobre o subsistema de comunicacao. pt_BR

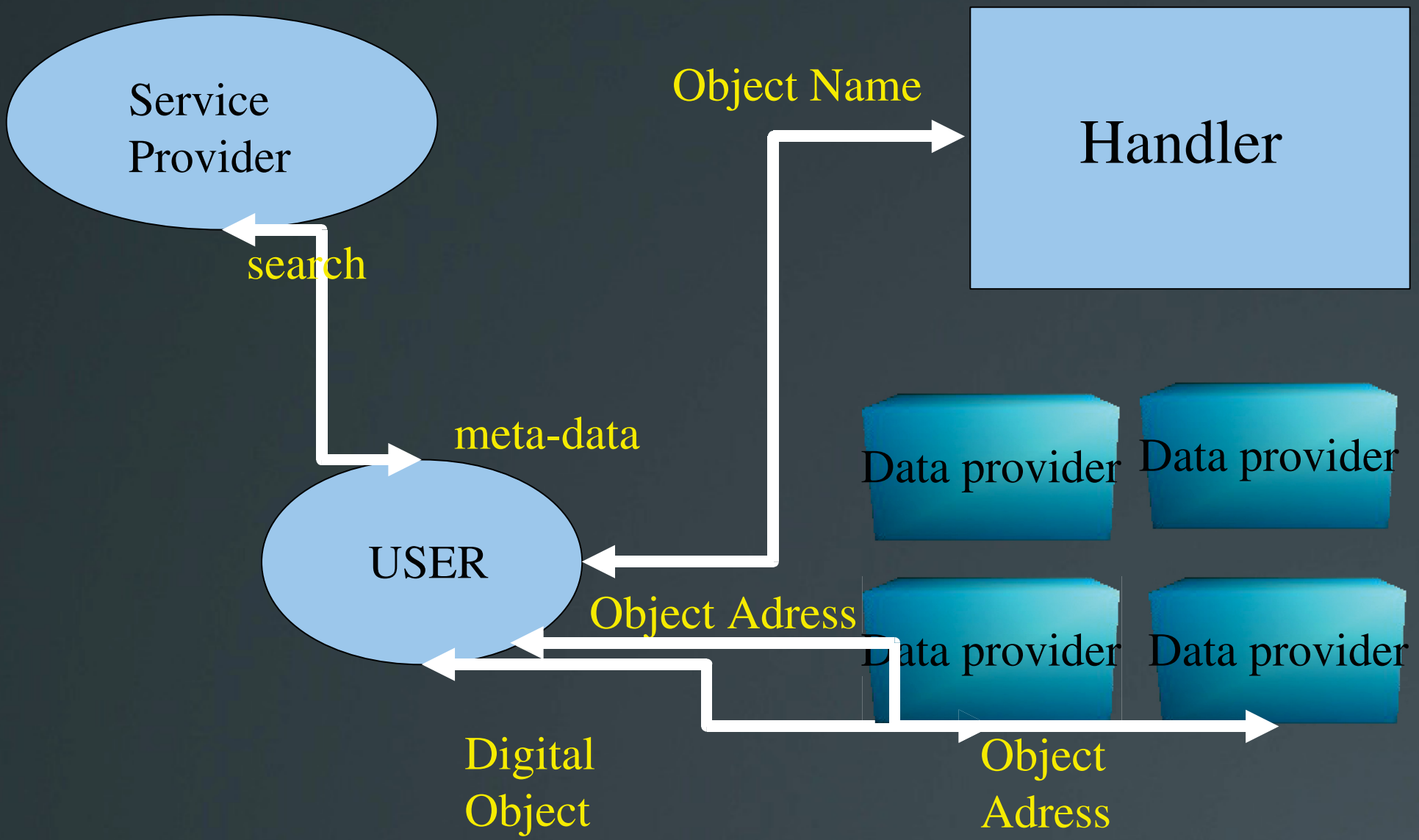
OAI-Dublin Core

- dc.format.mimetype text pt_BR
  - dc.language.iso po pt_BR
  - dc.source oai:ufpr.br:221072 pt_BR
  - dc.source.uri http://200.17.209.5:8000/cgi-bin/gw_42_13/chameleon.42.13a?host=localhost%201111%20DEFAULT&sessionid=VTLS&function=CARDSCR&search=KEYWORD&pos=1&u1=12101&t1=221072 pt_BR
- dc.title Avaliacao de desempenho do protocolo IEEE 802.2-LLC n

Preservação Digital

- Extensões P2P para OAI (FeeeLib, UFPR)
  - Uso de DHT para referenciar os conteúdos
  - Adaptação dos mecanismos de mapeamento (Identidade/Conteúdo)
    - Handle System
  - Copiar Conteúdo e metadado na mesma rede cooperativa
  - Criar um mecanismo de controle de confiabilidade

OAI/P2P UFPR


Service Provider
Object Name
Handler
search
meta-data
Data provider
Data provider
USER
Object Adress
Data provider
Data provider
Digital Object
Object Adress

Arquivamento Digital

- OAI/P2P
  - Preservação dos meta-dados existentes
  - Preservação da identidade
  - Maior Disponibilidade
  - Melhoria do desempenho
  - Autonomia (controle de conteúdo)
  - Agrupamento das informações
  - Busca global

Arquivamento 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)

Arquivamento 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/

Referências

- 5o Workshop de Redes Dinâmicas e Sistemas P2P Arquivamento Digital a Longo Prazo Baseado em Seleção de Repositórios em Redes Peer-to-Peer Tiago Vignatti1 , André Vignatti2 , Luis C. E. de Bona1 , Marcos Sunye
- Long-term Digital Archiving Based on Selection of Repositories Over P2P Networks IEEE p2p 2009