SCAPE

Ambientes de preservação escaláveis

4º Seminário sobre informação na Internet // Preservação digital

21 Nov. 2012

Miguel Ferreira, PhD.
Director técnico // KEEP SOLUTIONS
mferreira@keep.pt

KEEP SOLUTIONS
University of Minho SPIN-OFF

O que é o SCAPE?

Projecto de I&D em preservação digital

Visa o planeamento e execução de processos de preservação em grandes colecções de dados

Resultados esperados

- Definição de cenários de preservação (para guiar os desenvolvimentos e validar os resultados do projecto)
- Desenvolvimento de software/ferramentas open-source de auxílio à preservação digital
- Criação de workflows de preservação
- Edificação de um infra-estrutura de suporte à preservação (escalável)
- Criação de Guias de boas-práticas em preservação (migração de repositórios e preservação de dados científicos)

O SCAPE é uma continuação de um anterior projecto chamado PLANETS

KEEPSOLUTIONS
University of Minho SPIN-OFF

Financiamento

7º Programa Quadro (instrumento europeu de financiamento de investigação)

Projecto Integrado (IP) // Tecnologias da Informação e Comunicação (TIC)

Call sobre "Bibliotecas digitais e preservação digital"

Duração

42 meses // 3.5 anos

Fev 2011 a Jul 2014

Orçamento

11.3 Milhões de euros // 14.3 Milhões de dollars // ~30 milhões de reais

76% financiado pela Comissão Europeia

KEEPSOLUTIONS
University of Minho SPIN-OFF

KEEPSOLUTIONS

Institutos & fundações

AIT Austrian Institute of Technology

Internet Memory Foundation

Open Planets Foundation

Science & Technology Facilities Council

Universidades

UPMC Sorbonne Universités

TU berlin

TU Wien — Technische Universität Wien — Vienna University of Technology

The University of Manchester — Manchester 1824

Empresas

ExLibris — The bridge to knowledge

Microsoft Research

FIZ Karlsruhe — Leibniz Institute for Information Infrastructure

KEEPSOLUTIONS — University of Minho SPIN-OFF

Bibliotecas

British Library

Koninklijke Bibliotheek — National Library of the Netherlands

Österreichische Nationalbibliothek

Statsbiblioteket

KEEPSOLUTIONS — University of Minho SPIN-OFF

Detalhes

KEEPSOLUTIONS
University of Minho SPIN-OFF

O SCAPE irá estender o estado da arte através da

Criação de uma infra-estrutura que permita a execução de acções de preservação em larga-escala (na ordem dos milhões de ficheiros e vários TB de informação)

Controlo de qualidade automático

Integração com planeamento de preservação suportado por políticas e monitorização automática do meio ambiente

Validação do projecto

O projecto será validado recorrendo a 3 bancos de ensaio

1. Repositórios digitais oriundos da comunidade das bibliotecas
2. Conteúdos da Web oriundos de 3 arquivos da Web
3. Dados científicos oriundos da comunidade científica

Open-source

Suportado quase exclusivamente por tecnologias open-source

Resultados serão publicados com licenças APACHE 2.0

KEEPSOLUTIONS
University of Minho SPIN-OFF

Dados

- Repositórios digitais
- Conteúdo da Web
- Dados científicos

KEEPSOLUTIONS
University of Minho SPIN-OFF

Dados

- Repositórios digitais
- Conteúdo da Web
- Dados científicos

→ **Caracterização**
(Identificação de formatos, extracção de propriedades)

KEEPSOLUTIONS
University of Minho SPIN-OFF

Monitorização do ambiente
(Interno e externo)

Dados

- Repositórios digitais
- Conteúdo da Web
- Dados científicos

Caracterização
(Identificação de formatos, extracção de propriedades)

KEEPSOLUTIONS
University of Minho SPIN-OFF

Planeamento
(decisão sobre que acção tomar)

Monitorização do ambiente
(Interno e externo)

Caracterização
(Identificação de formatos, extracção de propriedades)

Dados

- Repositórios digitais
- Conteúdo da Web
- Dados científicos

KEEPSOLUTIONS
University of Minho SPIN-OFF

Políticas
(condicionam o planeamento)

Planeamento
(decisão sobre que acção tomar)

Monitorização do ambiente
(Interno e externo)

Dados

- Repositórios digitais
- Conteúdo da Web
- Dados científicos

Caracterização
(Identificação de formatos, extracção de propriedades)

KEEPSOLUTIONS
University of Minho SPIN-OFF

Políticas
(condicionam o planeamento)

Acções de preservação
(Migração de formatos)

Planeamento
(decisão sobre que acção tomar)

Monitorização do ambiente
(Interno e externo)

Dados

- Repositórios digitais
- Conteúdo da Web
- Dados científicos

Caracterização
(Identificação de formatos, extracção de propriedades)

KEEPSOLUTIONS
University of Minho SPIN-OFF

Políticas
(condicionam o planeamento)

Planeamento
(decisão sobre que acção tomar)

Acções de preservação
(Migração de formatos)

Controlo de qualidade
(Comparação de propriedades do objecto original com o objecto convertido)

Monitorização do ambiente
(Interno e externo)

Dados

- Repositórios digitais
- Conteúdo da Web
- Dados científicos

Caracterização
(Identificação de formatos, extracção de propriedades)

KEEPSOLUTIONS
University of Minho SPIN-OFF

Políticas
(condicionam o planeamento)

Planeamento
(decisão sobre que acção tomar)

Acções de preservação
(Migração de formatos)

Controlo de qualidade
(Comparação de propriedades do objecto original com o objecto convertido)

Monitorização do ambiente
(Interno e externo)

Dados

- Repositórios digitais
- Conteúdo da Web
- Dados científicos

Caracterização
(Identificação de formatos, extracção de propriedades)

KEEPSOLUTIONS
University of Minho SPIN-OFF

Componentes de preservação

Acções de preservação
(Migração de formatos)

Controlo de qualidade
(Comparação de propriedades do objecto original com o objecto convertido)

Políticas
(condicionam o planeamento)

Planeamento
(decisão sobre que acção tomar)

Monitorização do ambiente
(Interno e externo)

Dados

Repositórios digitais

Conteúdo da Web

Dados científicos

Caracterização
(Identificação de formatos, extracção de propriedades)

Componentes de preservação

KEEPSOLUTIONS
University of Minho SPIN-OFF

Planeamento e monitorização

Políticas
(condicionam o planeamento)

Planeamento
(decisão sobre que acção tomar)

Monitorização do ambiente
(Interno e externo)

Componentes de preservação

Acções de preservação
(Migração de formatos)

Controlo de qualidade
(Comparação de propriedades do objecto original com o objecto convertido)

Dados

Repositórios digitais

Conteúdo da Web

Dados científicos

Caracterização
(Identificação de formatos, extracção de propriedades)

Componentes de preservação

KEEPSOLUTIONS
University of Minho SPIN-OFF

Plataforma

Políticas
(condicionam o planeamento)

Planeamento
(decisão sobre que acção tomar)

Acções de preservação
(Migração de formatos)

Controlo de qualidade
(Comparação de propriedades do objecto original com o objecto convertido)

Monitorização do ambiente
(Interno e externo)

Repositórios digitais

Conteúdo da Web

Dados científicos

Caracterização
(Identificação de formatos, extracção de propriedades)

KEEPSOLUTIONS
University of Minho SPIN-OFF

Políticas
(condicionam o planeamento)

Acções de preservação
(Migração de formatos)

Planeamento
(decisão sobre que acção tomar)

Controlo de qualidade
(Comparação de propriedades do objecto original com o objecto convertido)

Monitorização do ambiente
(Interno e externo)

Gestão
técnica, projecto, científica

Adopção
disseminação, formação, sustentabilidade

Repositórios digitais

Conteúdo da Web

Dados científicos

Caracterização
(Identificação de formatos, extracção de propriedades)

KEEPSOLUTIONS
University of Minho SPIN-OFF

Dados e cenários de aplicação

KEEPSOLUTIONS
University of Minho SPIN-OFF

3 bancos de ensaio

Repositórios digitais

Dados científicos

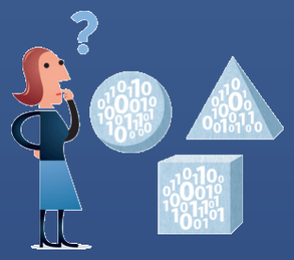
Arquivos da Web

Migração de TIFF para JPEG 2000

Redução de custos de armazenamento

80 TB // 2.2 M de páginas de jornais digitalizados do século XIX

Conversão de RAW para NeXus



Instrumentos de medição produzem dados em formato RAW. Pretende-se migrar todos esses dados para o formato NeXus baseado em XML por ser um standard internacional

Capacidade de processar ficheiros na ordem dos 100 GB

Controlo de qualidade em arquivos da Web

Conhecer o conteúdo do arquivo da Web para saber se é necessário agir para preservar

Melhorar os processos de recolha da Web de modo a garantir que não ficaram partes importantes por recolher

KEEPSOLUTIONS
University of Minho SPIN-OFF

Componentes de preservação

KEEPSOLUTIONS
University of Minho SPIN-OFF

Ferramentas de caracterização

Visam dar a conhecer os nossos dados

Identificação de formatos

Extracção de propriedades internas dos objectos

Nº de páginas, tamanho, largura, altura, esquema de cores, metadados, etc.

Exemplos: FITS, Tika, ffproble, Unix File, Droid, ...

Acções de preservação

Visam tornar os dados mais acessíveis

Migração de formatos

Emulação

Exemplos: ImageMagick, MS Sharepoint services, ffmpeg, gstreamer, ARC2WARC

Controlo de qualidade

Visam garantir que as acções de preservação não comprometeram a qualidade dos objectos

Comparam propriedades dos objectos antes e depois das acções de preservação terem sido aplicadas

Exemplos: Image Comparison, Matchbox, MarcAlizer VIPS, Jpylyzer

KEEPSOLUTIONS
University of Minho SPIN-OFF

Digital Preservation Toolkit

http://tinyurl.com/dptoolkit

KEEPSOLUTIONS
University of Minho SPIN-OFF

Planeamento e monitorização

KEEPSOLUTIONS
University of Minho SPIN-OFF

Monitorização do meio ambiente

Visa detectar eventos internos e externos ao nosso repositório de informação que possam constituir riscos ou oportunidades relevantes para a preservação dessa informação

Suporte para diferentes fontes de informação

- directórios de formatos, directórios de software, repositórios, utilizadores, tendências, etc.

Permite exprimir riscos e oportunidades sob a forma de "questões"

Notifica os interessados sempre que o estado do "mundo" viola uma determinada condição

Planeamento de preservação

Avaliação das diferentes alternativas de preservação com vista a maximizar os objectivos da instituição

Plato - http://www.ifs.tuwien.ac.at/dp/plato

Políticas institucionais

Definição formal de políticas institucionais que poderão influenciar decisões ao nível do planeamento de preservação

- Todos os dados devem ser preservados. Nada é eliminado
- O custo de uma ferramenta de preservação não deve exceder X
- O custo do salário de um especialista em preservação não deve exceder Y
- Todos os conteúdos devem ter um checksum

Plataforma

KEEPSOLUTIONS
University of Minho SPIN-OFF

Infra-estrutura onde todos os componentes são executados

Cluster Hadoop

Replicação de dados

Integração com repositórios

HDFS, HBASE, HDFS-Fedora

Três níveis de paralelização

Distribuição dos dados pelos vários nós do cluster (acesso mais rápido)

Divisão de grandes ficheiros para processamento paralelo

Paralelização de algoritmos e/ou do trabalho

Adopção

KEEPSOLUTIONS
University of Minho SPIN-OFF

Disseminação

- Sítio Web, Newsletter
- Twitter
- OPF Blog, Slideshare, Vimeo
- Github + OPF wiki
- Publicações científicas (22 artigos já publicados)
- Materiais promocionais (autocolantes, brochuras, pins)

Formação

- Várias formações previstas ao longo do projecto
- A primeira será nos dias 6-7 em Guimarães, Portugal (INSCRIÇÕES ABERTAS)

Sustentabilidade

- Evangelização através da participação em eventos
- Integração de resultados em aplicações comerciais
- Promoção do uso das aplicações

KEEPSOLUTIONS
University of Minho SPIN-OFF

Mais info

KEEPSOLUTIONS
University of Minho SPIN-OFF

Web Site

http://www.scape-project.eu

Código fonte

https://github.com/openplanets/scape/

Wiki de desenvolvimento

http://wiki.opf-labs.org/display/SP/Home

Workflows experimentais

http://www.myexperiment.org/search?query=SCAPE&type=all&commit=Search

Publicações

http://www.scape-project.eu/category/publication

Relatórios de projecto

http://www.scape-project.eu/category/deliverable

KEEPSOLUTIONS
University of Minho SPIN-OFF

KEEPSOLUTIONS

University of Minho SPIN-OFF

www.keep.pt

Miguel Ferreira
mferreira@keep.pt

Acknowledgements
Miguel Arellano (IBICT)
Dr. Ross King (AIT)

KEEPSOLUTIONS
University of Minho SPIN-OFF