Access IT Training

# Part 3.
# How to create a digital repository?

# Digital repository

- The digital repository is a place in which you deposit digital objects described with metadata
  - The outcomes of your digitisation projects
  - Any other suitable digital content
- You should consider issues related with depositing objects in preservation and presentation forms
  - Do you want to have two separate digital repositories or just one with all types of forms
- In order to create the digital repository you need at least three elements
  - Trained staff
  - Hardware infrastructure
  - Software infrastructure

# Digital repository

- Trained staff
  - The creation of a repository is one issue
    - It can be easily outsourced
    - It can be even done by company digitising your collections
    - It is one time process
  - The second and much more important one is the long term sustainability of the repository
    - It is a permanent process and it should be treated as such, especially in memory institutions
    - It causes continuous costs and all outsourcing aspects should be deeply analysed

# Digital repository

- Trained staff
  - What hardware and software infrastructure should be used
    - The one which is the best known by your staff
      - But avoid "The Untouchables" ;-)
    - The one for which you can easily get support
    - The one which has the chance to be supported for a long time

# Hardware

- What should be considered when choosing hardware solutions
  - Computing power and operational memory
    - For all crucial functions of the repository, especially browsing, indexing and searching, content consistency check etc.
  - Storage size and type
    - On-line access vs long-term archiving
    - Speed vs capacity
  - Network bandwidth
    - For upload of digital objects
      - Only local network or wider cooperation?
    - For download of digital objects
      - Number of parallel sessions vs number of parallel requests
        - http://dlibra.psnc.pl/wbc/wbc-index.html
    - Summary bandwidth vs maximum transfer speed

# Hardware

- What should be considered when choosing hardware solutions
  - Scalability
    - What to do when the hardware capacity ends?
      - Replace old server?
      - Add another one?
    - Virtualization is a commonly used solution (when you have proper funding)
      - The real hardware infrastructure is separated from the virtual hardware infrastructure accessible for software
      - The digital library software should/must also address scalability issues
  - Example configurations:
    - http://dlibra.psnc.pl/community/pages/viewpage.action?pageId=6258894

# Hardware

- Finally try to consult other institutions which can have required experience
  - As long as these are not commercial companies, there is a big chance to get useful and unique information

# Software

- There are several software platforms
  - Free
    - DSpace
    - Fedora (in fact not for use without additional tools)
    - Greenstone
    - ePrints
    - CDS Ingenio
    - …
  - Paid
    - Expensive
      - VTLS Vital (based on Fedora)
      - ExLibris DigiTool
    - Cheap (but still good)
      - PSNC dLibra ;-)
      - …

# Areas of software functionality to take into account

- **Content Management** – Tools and procedures that support upload of content to the digital library and control of this process of submitting and versioning
- **Content Acquisition** – Import and export of content, support of various formats of documents
- **Metadata** – Support of various metadata formats is important for indexing, upload of content, making it accessible and content protection

# Areas of software functionality to take into account

- **Search Support** – It applies to numerous searching and browsing functions, search in the metadata, full text search, hierarchical browsing and so on.
- **User Management and Privacy Protection** – User management and privacy protection includes administration of passwords, user accounts with access rights with the possibility to retrieve forgotten passwords and so on.

# Areas of software functionality to take into account

- **Support of Reports and Queries** – This criterion deals with evaluation of the digital library and possibility to monitor the patterns of users' behavior to improve the services provided and the usage of user activity logs for billing purposes.
- **Sustainability, Data Protection** – Protection of the metadata, consistency and integrity of the database, backup, possibility to support the migration of data.

# Areas of software functionality to take into account

- **Interoperability** – It allows two-way cooperation with other distributed systems on the level of the metadata, search and acquisition and providing of documents, OAI-PHM support
- **User Interface** – This category deals with the support of more languages and allows adapting of the user interface according to various needs of users or different implementations

# Areas of software functionality to take into account

- **Standards Compliance** – Standards are important for sharing and long-term storage of digital content.
- **Automation Tools** – This category deals with tools for automated acquisition of contents, harvesting, generating of the metadata, maintenance and so on.
- **Support, Services** – An important aspect of all software systems. Numerous good digital libraries come from the area of Open Source, where this aspect must be taken into account in particular. Important services are: documentation, helpdesk, collection of requirements to improve it, discussion forums and so on.

# Greenstone

- Year of creation: 1997
- Development organisation: University of Waikato (New Zealand)
- Programming language:
  - Ver. 2: PERL
  - Ver. 3: Java (research version of Greenstone)
- Database: Not neccessary (GDBM is used)
- Metadata format: Dublin Core
- http://www.greenstone.org/

# EPrints

- Year of creation: 2000
- Development organisation: University of Southampton (UK)
- Programming language: PERL
- Database: MySQL
- Metadata format: Dublin Core
- http://www.eprints.org/

# DSpace

- Year of creation: 2002
- Development organisation: DSpace Foundation (MIT/HP)
- Programming language: Java
- Database: PostgreSQL, Oracle
- Metadata format: Dublin Core
- http://dspace.org/

# Software

- EPrints and DSpace were (initially) developed in an academic environment with the focus on publishing the scholarly material in mind
- Greenstone 2 was also developed in an academic environment, but its development was supported by UNESCO and focused on use in developing countries
- Majority of digital library systems are described as systems suitable for any use, but the actual usability should be assessed in a context of particular project or institution

# Software

- Trainings for different OS software are often organised as tutorials connected with international DL conferences
- For popular tools there is also a lot of on-line documentation and courses available
  - For details check the website of these software systems

# What software to choose?

- Please analyse
  - Functionality
  - Technical requirements
  - Availability of documentation, support, services and trained staff
    - Especially the availability of local users community
- Think in the long-term perspective

# Practical part

- We want to show you very shortly a sample digital repository software created on the basis of dLibra system
  - We think that there are some aspects of this software which may be interesting for you and which may be important when you will be setting-up a new digital repository using any software package