

***Lakes and rivers:  
Metadata management  
and the  
Magazzini digitali (Digital Stacks)  
prototype***

# Magazzini Digitali (Digital Stacks)

- ✓ The Digital stacks project (started at the end of 2006) aims
  - ✓ to set up a prototype of an “Open Archival Information System”
  - ✓ in order to ensure a long term viability, renderability, authenticity and availability of the acquired digital resources

# Digital Stacks: what's in a name

- ✓ The name of the project recalls the stacks of the legal deposit libraries
- ✓ Quoting an important European project on DP [NEDLIB = Networked European Deposit Libraries 1997-2000]:
  - ✓ “For us, as memory organizations, this means we have to move from paper-based stacks to digital stacks”

# Digital Stacks: digital and conventional

- ✓ In most aspects digital stacks are comparable to the conventional ones:
  - ✓ digital resources have to be preserved for the long term;
  - ✓ digital stacks grow by adding new resources: modification and deletion is not an option;
  - ✓ it is impossible to predict the usage frequency of stored digital resources: probably some resources will be seldom or never used

# Digital Stacks: underlying principles, 1

- ✓ Taking into account the fact that "component failures are the norm rather than the exception" the fault tolerance of the overall system is based on
  - ✓ data replication (on different machines located in different sites)
  - ✓ simple and widespread hardware components, non vendor-dependent, that can be easily replaced (just simple personal computers).
- ✓ Nowadays an ordinary personal computer could easily store up to 4 TB (ex. g. equipped with four 1000 GB hard disks ) using widespread and inexpensive SATA technology

# Digital Stacks: underlying principles, 3

- ✓ data replication relies on open source disk synchronization utility (rsync ): to avoid hardware dependencies (ex. g. disk controllers) RAID is not used
- ✓ a third site acting as a dark archive (for disaster recovery purposes) based on a different technology has been created as a provision to enhance the security of the overall system (i.e. a backup using LTO Ultrium3 tapes)

# Digital S

# Principles, 2

- ✓ The following
  - ✓ ten rack-mounted large capacity hard drives in accordance with the National Center for Digital Library
  - ✓ an open source operating system (Fedora distribution)
  - ✓ A working prototype by the end of 2004



set up:

equipped with four hard drives installed in accordance with principle (5 at the National Center for Digital Library)

operating system selected (Linux)

is available since

# Digital Stacks: who we are

- ✓ Biblioteca Nazionale Centrale di Firenze
- ✓ Biblioteca Nazionale Centrale di Roma
- ✓ Fondazione Rinascimento Digitale



Magazzini

Digitali



# The Digital Stacks core: two kinds of file (container)

- ✓ **Data** (wrapped in WARC container)
  - ✓ aggregate **digital objects** for ease of storage in a conventional file system.
- ✓ WARC = ISO/PRF 28500
  - ✓ PRF= proof

+

- ✓ **Metadata** (wrapped in XMLTAPE Container) aggregate XML records (also from different metadata Schemas) for ease of storage in a conventional file system.
- ✓ “An XMLtape is an XML file that concatenates the **XML-based representation of multiple digital objects**”
- ✓ XMLTAPE is based on aDORe project [Van De Sompel et al.]
- ✓ A **Long Term** Archive can not rely on **lake model** = stores of metadata based on few “Schemas” and fed by a few principal sources
- ✓ ]

# The Metadata management problem

- ✓ A **Long Term** Archive can not rely on **lake model** = stores of metadata based on few “Schemas” and fed by a few principal sources
- ✓ A **Long Term** Archive has to face stores of metadata based on “Schemas” that can change over time and which are fed by many streams. Probably the most suitable model is the **river model**

[lake and river = Eric Hellman, Lorcan Dempsey]

# The River model - 1

- ✓ In a Long Term archive it is realistic to assume that
  - ✓ different metadata Schemas originating from different “agents” [metadata harvesters OAI-PMH, Metadata extractors like JHOVE, Librarian, etc]
  - ✓ every Schema can evolve over time;
  - ✓ there could be some semantic overlap between elements belonging to different Schemas (ex. g. PREMIS, MIX, ecc)

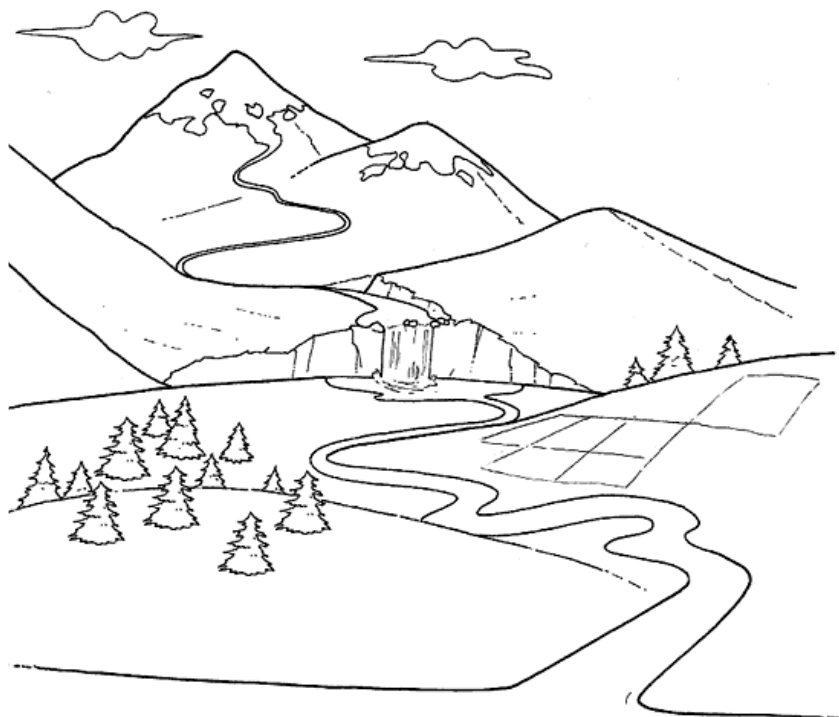
# River (not Babel) model

- ✓ Since
- ✓ Metadata are an essential mean to “manage” Data
- ✓ it is essential
- ✓ to have a “full control” of Metadata



# The river model: tools available?

- ✓ No tools available for a coordinated management of different Schemas / formats
- ✓ Some directions we are considering:
  - ✓ Crosswalks like MORFROM (demonstration OCLC web service, limited to bibliographic metadata )
  - ✓ Dspace future plans “HP and MIT also have a research project called SIMILE that is investigating how to support arbitrary metadata schemas using RDF”
  - ✓ It's important to take into account the PREMIS concept of **semantic unit** (piece of information the archive need to know) vs **metadata element** (an implementation decision)



***Grazie***

**谢谢**