

# Magazzini Digitali (Digital Stacks) turning a current prototype into an operational service

Giovanni Bergamin

(Biblioteca Nazionale Centrale Firenze)

Maurizio Messina

(Biblioteca Nazionale Marciana – Venezia)

# Outline

- ✓ From prototype to service: a look to the *Digital stacks* technical architecture
- ✓ Legal and agreements framework:
  - ✓ An agreement with electronic publishers for access to and use of legal deposit documents
  - ✓ An agreement between the deposit libraries

# Magazzini Digitali (Digital Stacks): not only a technical project

- ✓ a) economic implications
- ✓ b) selection problems
- ✓ c) a legal aspects (access and use);
- ✓ d) cooperation between legal deposit institutions
- ✓ ...

# Magazzini Digitali (Digital Stacks): a definition of digital preservation

- ✓ Digital preservation could be defined as
- ✓ a service to be provided by trusted digital repositories
- ✓ in order to ensure – for deposited digital resources –
  - ✓ viability,
  - ✓ renderability,
  - ✓ authenticity (identity + integrity)
  - ✓ and availability
- ✓ for designated communities.

# 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 conventional ones:
  - ✓ digital resources must be preserved for the long term;
  - ✓ digital stacks grow as new resources are added;
  - ✓ modification and deletion is not an option;
  - ✓ it is impossible to predict the usage frequency of stored digital resources;
  - ✓ and it is likely that some resources will be seldom or never be used

## Digital stacks: a vanity search ...

- ✓ This expression is used also in different projects within the context of digital preservation
- ✓ Ex. g. “Digital stacks: rather than boxes, shelves, and climate controlled environments, digital information must be stored in containers, file systems, and secure servers”.

# Digital Stacks: underlying principles, 1

- ✓ the aim of the project was to set up an infrastructure based on a "long term framework".
- ✓ taking into account the fact that component failures are the norm rather than the exception, the infrastructure is based on
  - ✓ data replication (different machines located in different sites)
  - ✓ simple and widespread hardware components, non vendor-dependent, that can easily be replaced (just simple personal computers).

## Digital Stacks: underlying principles, 2

- ✓ The infrastructure does not rely on custom or proprietary software but is based on an open source operating system and utilities (widespread acceptance means less dependencies).
- ✓ Now an ordinary personal computer could easily store up to 8 TB (equipped with four 2000 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

# Turning a current prototype into a operational service

- ✓ the original plan was to use an offline storage system (ex. g. LTO tapes) to set up the *dark archive* for disaster recovery purposes
- ✓ ... but for the operational service we decided to use the same technology used in the two *light archives* (i. e. *online (\*)* storage using just simple personal computers ).
- ✓ (\*) the use of the term *online* here does not change the purpose of the *dark archive* that is “to function as a repository for information that can be used as a fail-safe during disaster recovery”

## Turning a current prototype into a operational service – 2

- ✓ LTO is a robust and reliable solution but introduces technology dependencies (ex. g. "robots") and media management problems.
- ✓ For the same reasons we decided not to use an HSM (Hierarchical storage management) system (there are different implementation based on proprietary systems).

# Online vs Offline storage

- ✓ Comparing all the costs of online and offline storage is not an easy task.
- ✓ Ex. g.:
  - ✓ the cost of SATA disks is decreasing day by day while their capacity is increasing,
  - ✓ but it is difficult to estimate the so called total cost of ownership of a tape based solution.
- ✓ Taking into account all the pros and cons, we concluded that the most convenient solution is online storage on simple and easily replaceable personal computers (“easily replaceable” means replaceable with no or minor impact on the overall architecture).

# Online storage ecology

- ✓ the power consumption of the storage computers (and the carbon dioxide emissions) is a problem
- ✓ However in the last years:
  - ✓ the “green computer” technology (i. e. more energy-efficient versions of computers) is gaining widespread market awareness.
  - ✓ Solid State Drive (SSD) is a rapid developing technology (this solution could significantly reduce in the near future the energy consumption)

# Turning a current prototype into a operational service, 3

- ✓ The current Digital stacks prototype is now turning into an operational service based on
  - ✓ two main deposit sites (managed by the *Biblioteca Nazionale Centrale di Firenze* and by the *Biblioteca Nazionale Centrale di Roma*)
  - ✓ a dark archive (managed by the *Biblioteca Nazionale Marciana, Venezia*).
- ✓ The *Fondazione Rinascimento Digitale* will continue to support and promote the Digital stacks operational service.



Magazzini

Digitali

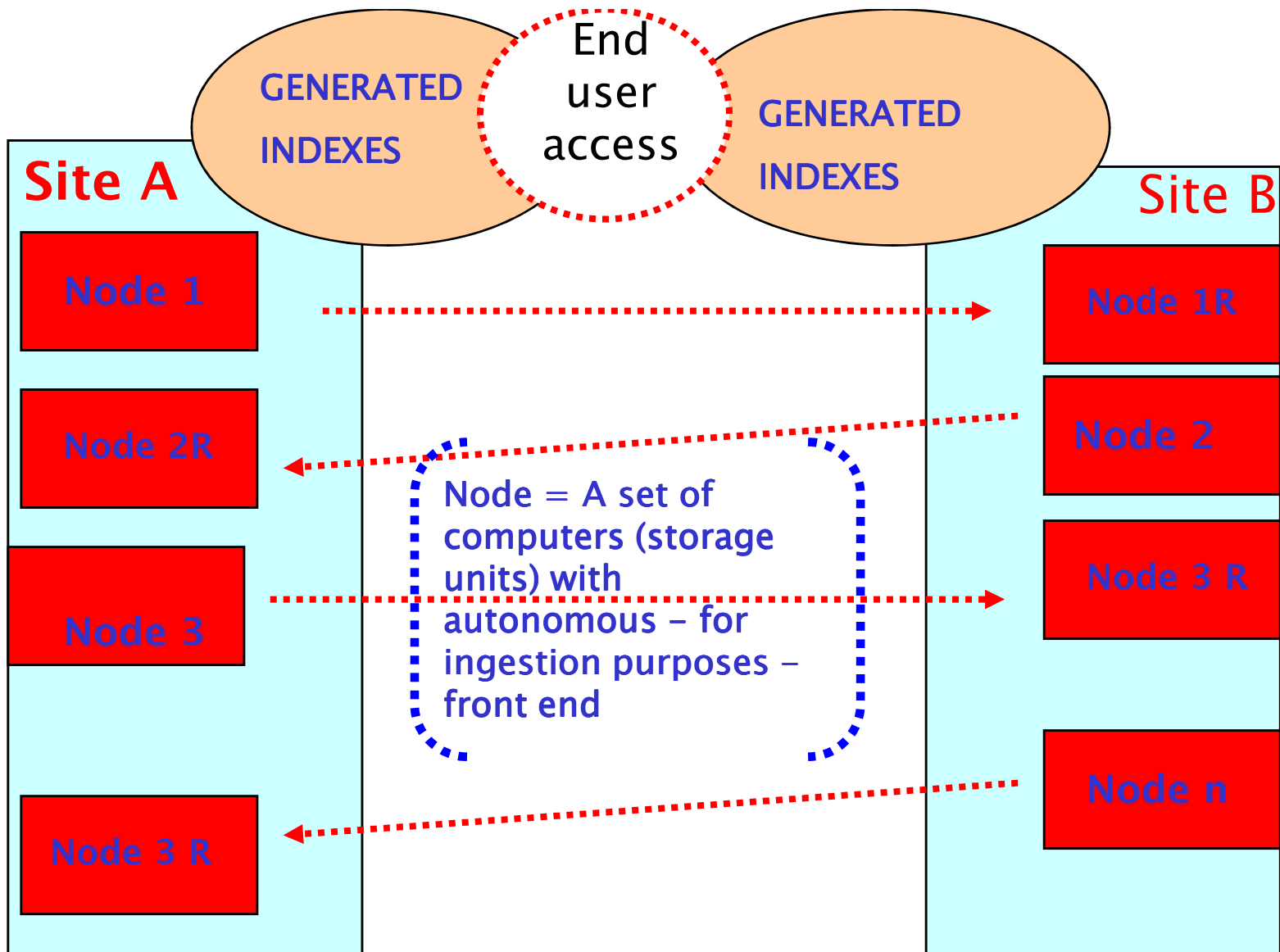


# Digital stacks architecture – 1

- ✓ every main *site* is composed by a set of autonomous and independent *nodes*.
- ✓ every *node* on a given site has a mirror node on the other site
- ✓ digital stacks service does not rely on “*master site / mirror site*” architecture and every site will contain – in a symmetrical way – both *master nodes and mirror nodes*.

# Digital stacks architecture – 2

- ✓ Every physical file is replicated 2 times on different computer within the same node
- ✓ Dark archive contains also two copies of this file on two different computer.
- ✓ As a result within Digital stacks every physical file is replicated 6 times.



**DARK ARCHIVE**

# Digital stacks architecture – 3

- ✓ Setting up one main site in Florence close to the Arno river and the Dark Archive in Venice with the well known "acqua alta" (or high tide) problem, could result in a relevant threat for the security of the overall service.
- ✓ One important decision was to locate all the hardware on external data centers (aka collocation centers ).

# Digital stacks architecture – 4

- ✓ Certification to ISO 27001 international security standard, will be the basic prerequisite for the selection of a data center.
- ✓ Every institution (Florence, Rome and Venice) will select 3 different data centers owned and managed by 3 different companies (to reduce the risk of “domino” effects).

# Digital stacks architecture – 5

- ✓ the 3 collocation centers have to be distant from each other by at least 200 km (to reduce the risk of natural threats).
- ✓ This architecture based on certification to ISO 27001 international security standard will form the basis for a domain specific certification of Digital stacks as trusted digital repository
  - ✓ (during the prototype phase we tried to apply DRAMBORA but also TRAC was taken into account).

# Digital stacks core

- ✓ Digital Stacks could ingest two kinds of file:
  - ✓ *data* wrapped in WARC containers: WARC (ISO 28500) container aggregates digital objects for ease of storage in a conventional file system.
  - ✓ *metadata* wrapped in MPEG21-DIDL containers : MPEG21-DIDL (ISO 21000) is a simple and agnostic container suitable for the representation of digital resources (sets of metadata conformant to different *Schemas*)

# The Metadata management problem

- ✓ A Long Term Archive can not rely on lake model  
= stores of metadata based on one(?) or 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. It  
could be based only on a 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 *Schema* originating from different “agents” [metadata harvesters OAI-PMH, Metadata extractors like JHOVE, Librarians, etc]
  - ✓ every *Schema* can change over time;
  - ✓ there could be some semantic overlap between elements belonging to different *Schemas*
- ✓ “Schemas express shared vocabularies and allow machines to carry out rules made by people”: <http://www.w3.org/XML/Schema>

# River (not Babel) model

- ✓ Since Metadata are an essential mean to “control” Data
- ✓ In a Long Term Archive it is essential to “control” Metadata to avoid the risk of a Babel model



# The river model: tools available?

- ✓ No tools available for a coordinated management of different Schemas / formats
- ✓ Some directions:
  - ✓ 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”

# Legal and Agreements Framework – 1

- ✓ The *Commitment*: L. 106/2004 – D.P.R. 252/2006 (art. 37, comma 2): a trial period for the legal deposit on a voluntary basis of electronic documents, that are defined by the law as “documents disseminated via a computer net”.
- ✓ Funded by MiBAC, General Direction for Libraries, with the support, in terms of human and financial resources, of Fondazione Rinascimento Digitale (FRD)
- ✓ BNCF, BNCR, BNM, FRD

# Legal and Agreements Framework – 2

## Trial period goals

- ✓ To implement an *organizational* model: national and regional archives of electronic publishing production
- ✓ To implement a *service* model: balancing the right-holders interests in contents protection with the final users ones in contents access
- ✓ To implement a *long term preservation system*: long term preservation and access to digital contents, as well as their authenticity (identity and integrity)

# Legal and Agreements Framework – 3

An agreement between all the involved stakeholders

- ✓ to define specific roles and responsibilities of each institution from different points of view: scientific, technical, operational and financial
- ✓ to set up a steering committee for all management, monitoring and results assessment activities
- ✓ to define an organizational and financial sustainability plan

# Legal and Agreements Framework – 4

An agreement with electronic publishers for access to and use of legal deposit documents

- ✓ Documents harvesting
- ✓ Clearances in case of license subject documents; file formats TBD
- ✓ 2 copies each in BNCF and BNCR, 2 off-line copies in BNM
- ✓ ISO 27001 certified Data Centers; ISO 14721 OAIS Digital Archives
- ✓ Changes tracking, Long term preservation actions allowed
- ✓ Registered users access to the documents on the libraries LANs
- ✓ Specific agreements for files printing and downloading, with compensation system for right-holders if necessary
- ✓ Allowed access in regional deposit libraries LANs, but only to documents of publishers who are in the same region of the deposit library

# Legal and Agreements Framework – 5

## Extending the test basis?

- ✓ Legal deposit born digital resources, i.e. e-journals, and also Ph. D. digital thesis, resulting from specific agreements with universities
- ✓ Digital resources resulting from digitisation projects funded by the Italian Digital Library initiative, mainly in the memory institutions range and only for master copies

# Legal and Agreements' Framework – 6

A way (not the only one!) for Sustainability

- ✓ An agreement with publishers to fulfill the *perpetual access* provision of e-journals licenses, through Trusted Digital Repositories managed from the legal deposit libraries network