Digital Stacks

The Italian Digital Preservation Service

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Giovanni Bergamin
(Biblioteca Nazionale Centrale di Firenze)
Maurizio Messina
(Biblioteca Nazionale Marciana - Venezia)

Ankara – 2011.07.07

Outline

- ***Digital Stacks*** aims to set up a long term digital preservation system for electronic documents published in Italy and distributed via digital communication network, according to the legal deposit law (L. 106/2004, DPR 252/2006)

- This Presentation:
  - technical architecture
  - metadata
  - legal and agreements framework
  - organizational and service model
  - sustainability

Digital Stacks: not only a technical project

- legal aspects (copyright)
- economic implications (sustainability)
- cooperation between legal deposit institutions
- materials selection and appraisal
- …

The Digital Stacks definition of digital preservation

a public service to be provided by trusted digital repositories in order to ensure - for deposited digital resources -

- Viability (i.e.: permanence over the long time of a bit sequence)
- Renderability (i.e.: property of the bit sequence to be readed from a device in order to be displaied to a user)
- authenticity (i.e: identity + integrity)
- availability

for designated communities

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 used

Digital Stacks: underlying principles, 1

to set up an infrastructure based on a "long term framework":

- 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: nowadays 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, 2

open source operating system and utilities
(widespread acceptance means less dependencies)

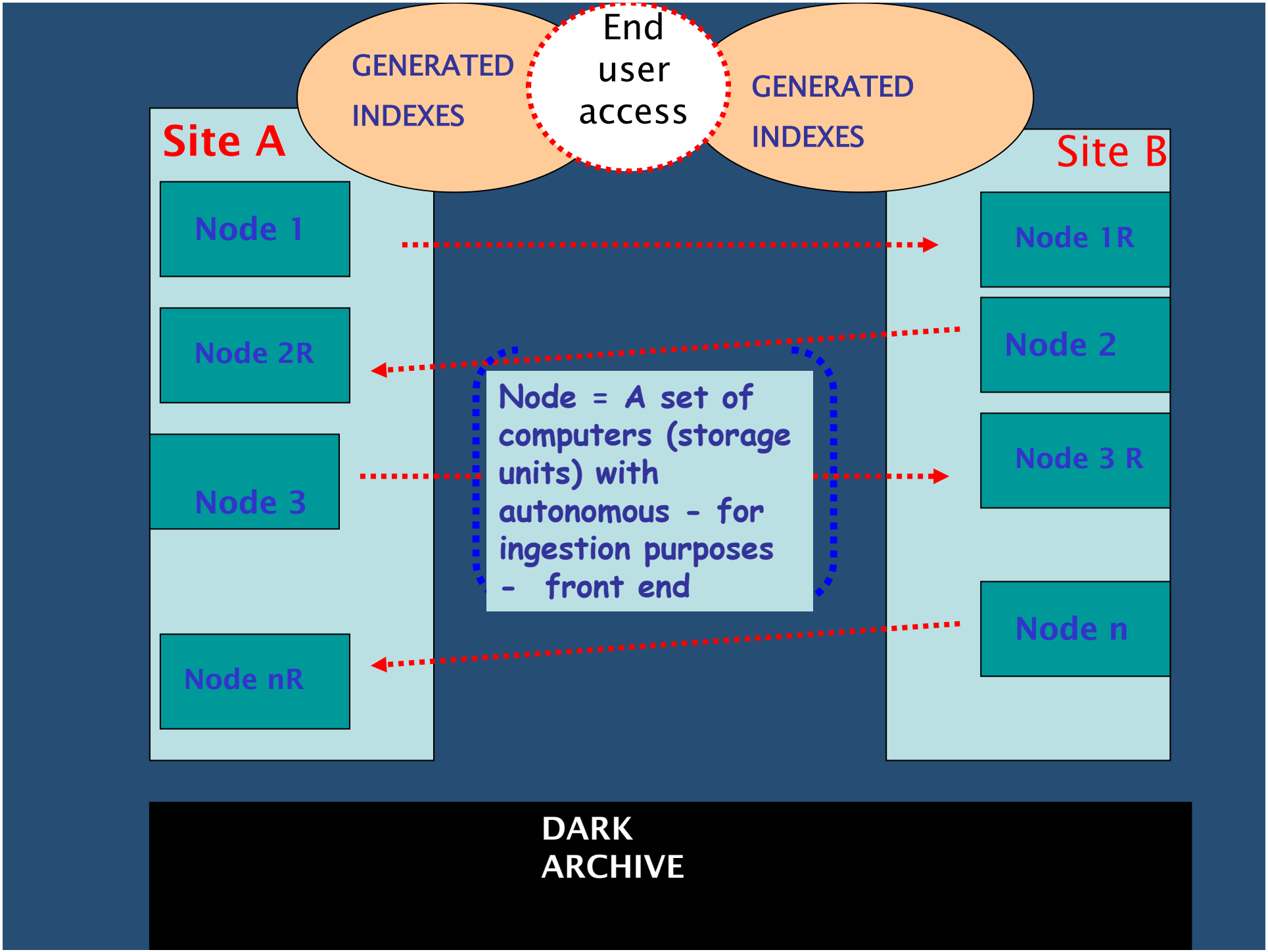
- data replication relies on open source disk synchronization utility (rsync for UNIX)
- to avoid hardware dependencies (ex. g. disk controllers) RAID is not used

The Digital Stacks system

- the Digital Stacks system is based on
  - two **main** deposit **sites** (managed by the *Biblioteca Nazionale Centrale di Firenze* and by the *Biblioteca Nazionale Centrale di Roma:* services and preservation)
  - a **dark archive** (managed by the *Biblioteca Nazionale Marciana,* Venezia: preservation only).
- The *Fondazione Rinascimento Digitale* keeps on supporting and promoting the Digital stacks operational service.

Digital Stacks architecture - 1

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


GENERATED INDEXES
End user access
GENERATED INDEXES
Site A
Site B
Node 1
Node 2R
Node 3
Node nR
Node 1R
Node 2
Node 3 R
Node n
Node = A set of computers (storage units) with autonomous - for ingestion purposes - front end
DARK ARCHIVE

Digital Stacks architecture - 2

- each physical file is replicated twice on different computers within the same node
- the *Dark archive* contains also two copies of the same file on two different computers
- as a result within Digital Stacks each physical file is replicated 6 times

The *Dark Archive*, 1

- 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"

The *Dark Archive*, 2

- LTO is a robust and reliable solution but introduces technology dependencies (ex. g. "robots") and media management problems
- as regards costs, the cost of SATA disks is decreasing day by day while their capacity is increasing
- it is difficult to estimate the so called total cost of ownership of a tape based solution
- for the same reasons we decided not to use an HSM (Hierarchical storage management) system (there are different implementations based on proprietary systems)

Digital Stacks Repositories -1

- setting up one main site in Florence close to the Arno river (flood risks!) 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 Repositories - 2

- ISO 27001 international security standard certification is the basic prerequisite for the selection of a data center
- each institution (Florence, Rome and Venice) will select 3 different data centers owned and managed by 3 different companies (to reduce the commercial risk of "domino" effects)
- the three *collocation centers* have to be at least 200 km far away from each other (to reduce the risk of natural threats).

Digital Stacks Repositories - 3

- This architecture, based on compliance certification to ISO 27001 international security standard, is 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 repositories will also be ISO 14721 - OAIS compliant

Digital Stacks: data and metadata

Digital Stacks can 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 compliant to different *Schemas*)

The Metadata management problem

- a **Long Term** Archive can not rely on a **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

In a Long Term archive we assume that:

- There will be different metadata *Schemas* originating from different "agents" [metadata harvesters OAI-PMH, Metadata extractors like JHOVE, Librarians, etc]
- each *Schema* can change over time
- there could be some semantic overlap between elements belonging to different *Schemas*

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 Tower 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)
  - ***Linked Data:*** a set of best practices for publishing and connecting structured data on the Web, based on: URIs (to identify things), HTTP URIs (so that these things can be referred to and looked up by people and user agents), RDF/XML (to provide informations about these things)

Legal and Agreements Framework - 1

- the *Commitment:* L. 106/2004 - D.P.R. 252/2006 (art. 37): a trial period for the legal deposit on a voluntary basis of electronic documents, that are defined by the law as "documents disseminated via digital communication network "

- funded by MiBAC, General Direction for Libraries, with the support, in terms of human and financial resources, of Fondazione Rinascimento Digitale (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 interests in accessing the contents
- 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 BNCF, BNCR, BNM, FRD

- 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

signed 2010-01-19

Legal and Agreements Framework - 4

An agreement with electronic publishers to fulfill law provisions

- documents harvesting
- clearances in case of license subjected documents; file formats t.b.d. (e.g. WARC)
- 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
- Files printing and/or downloading under a specific license
- allowed access in regional deposit libraries LANs, but only to documents of publishers who are in the same region of the deposit library

to be signed July 14th, 2011

Legal and Agreements Framework - 5

A license model for legal deposit documents printing and/or downloading, with different service levels:

- three different paper printing possibilities: 15%, 50%, or the whole document
- three different possibilities to send printed documents or files: only to italian legal deposit libraries; to all italian libraries; to all libraries everywhere
- two different ways of sending documents to libraries: postal service or fax for printed documents; professional services (e.g. ARIEL, NILDE, …) for files, with file destruction after the download
- two different downloading possibilities: only for registered users inside the legal deposit library premise; for registered users inside and outside the legal deposit library premise, in the last case via professional services (NILDE, ARIEL, …), with file destruction after the download
- choices are up to publishers
- no commercial utilization allowed
- no economic compensation for publishers for two years, t.b.d. after

Extending the test basis?- 1

www.depositolegale.it

- 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

Extending the test basis?- 2

Ph. D. Digital Thesis

- cooperation with CRUI *Open Access Group*
- harvesting of PH. D. Thesis deposited in the universities Open Archives
- harvesting of Ph. D. Thesis subject to "*embargo*", with access restricted to registered users on the libraries LANs, through workstations without periferal devices
- PDF(A), *Dublin Core*, MPEG21-DIDL
- *Eprints3*, *DSpace 1.5*
- Test underway, 17 universities harvested

Sustainability?

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

Digital Stacks: State of the Art

- Rome Data Center: awarded (Bologna), HW mounted, stress test completed
- Venice Data Center: awarded (Roma), HW awarded
- Florence Data Center: tender underway

(to be completed by September 2011)

The Digital Stacks Service will be working from January 2012

Digital Stacks: Next Steps

- printing/downloading management SW (according to Agreement and License provisions)
- SW engineering (SIP,AIP,DIP OAIS phases)
- audit and certification (as a trusted digital repository)
- system maintenance and evolution
- HW assistance and maintenance (after 3 years)
- sustainability!! (business model)