ina EXPERT

Enseignement
Formation
Recherche

Big Data, exploiter de grands volumes de données

mardi 3 juillet 2012

Daniel Teruggi, Head of Research
dteruggi@ina.fr
Ina: Institut National de l'Audiovisuel

Institut national de l'audiovisuel

ina EXPERT
Enseignement
Formation
Recherche

Missions:

- National Broadcast Archive, collections starting in 1933 for Radio and 1949 for Television
- Legal Deposit for Radio and Television since 1995
- Research and Experimentation (participation to major projects: PrestoSpace, Quaero, Caspar, PrestoPRIME, etc.)
- Archive based production
- Professional Training, Master in Multimedia Production and Digital archiving

Status:

- Public Industrial and Commercial Enterprise: EPIC
- 1000 employees
- 129M€ annual budget

1959 - Tournage "Cinq Colonnes à la Une" pendant la Guerre d'Algérie
André Hugues - Photo INA

Institut national de l'audiovisuel

ina EXPERT
Enseignement
Formation
Recherche

Professional Heritage Archives:

1 350 000 hours, 700 000 hours of radio and 650 000 hours of television
(Public broadcasters)
Ina owns the Production Rights, 30000 new hours every year.
Any possible analogue format!
Highly menaced collection, 850.000 hours already digitized and on-line

Radio and Television Legal Deposit:

3 000 000 hours, representing 100 TV channels 20 radio channels
(Public and Private)
800.000 new hours every year, digital media on shelf (for the moment).
No commercial use, access restricted for research only.

Legal Deposit of Internet since 2009 (6000 websites)

Ina makes this memory accessible

ina EXPERT
Enseignement
Formation
Recherche

To professionals – through digitization and theme organisation

ina média
ina's archive collections online

Since 2004 Ina proposes a unique service
www.inamediapro.com

Largest on-line collection of digitized archives, more than 850 000 hours of Images and almost 6 million documents
french and english versions
5 000 registered users, 35% coming from outside France

To scientists – to permit analysis and comprehension of the audiovisual domain

8 000 researchers are registered at the Inathèque premises at the National Library for access to the Legal Deposit

To the public – sharing the national memory

Since April 2006, through the website
www.ina.fr

Ina proposes an on-line access to 30 000 hours of radio and television, for streaming or downloading. Specific contents dedicated to educational issues for secondary schools

Some figures concerning audiovisual contents

1 year: 8760 hours

1 000 000 hours represent : 114 years of continuous visioning or listening

Digitizing 1 000 000 hours may roughly cost:

500 000 hours of Sound:	500 000 x 30€	= 15 000 000 €
300 000 hours of Video:	300 000 x 90€	= 27 000 000 €
200 000 hours of Film:	200 000 x 300€	= 60 000 000 €
Total:		**102 000 000 € (129 M$)**

100 000 hours at 50mB = 18 petabytes of data

Documenting 1 hour takes roughly 4 hours

1 000 000 hours = 4 000 000 hours = 456 years

Right clearance may take 1 day per hour (very optimistic)

Documentation is a major issue for Ina:

ina EXPERT
Enseignement
Formation
Recherche

- Most of the video and film contents are already hand-documented, including Right management issues
- Many old audio documents are badly documented
- Documentation is strongly use oriented, description schemes were inherited from broadcast production
- Documentation is based on several continuously updated Thesaurus
- Ina is evolving from a mass Archive to an Archive structured in collections and themes
- Ina will soon open its archives to user generated content with the project “Mémoire partagée”

Documentation evolution in the recent years:

ina EXPERT
Enseignement
Formation
Recherche

- 180 documentalists work in Ina to catalogue incoming contents, structure collections and update documentation
- Today there is an identified need for documentation and segmentation assistance, in order to accelerate hand-based documentation
- The passage from B2B to B2C has had strong implications in the documentation schemes and necessities
- The perception of documentation is changing, dynamic multimedia indexation will be a major issue in the future for professional and consumer uses
- A complete documentation doesn’t exist

Major challenges for indexation:

ina EXPERT
Enseignement
Formation
Recherche

- Indexation and documentation are never ending actions, they depend on uses and change objective through time
- Hand documentation is extremely precise and efficient for large collections, however it is not updatable and cannot deal with the amount of entering material
- Scalability is the major issue when searching for a content among millions, mainly when different contents are searched on the same document
- Organising contents in structured collections based on themes or other criteria
- Searching a specific content versus exploring a collection

Some essential questions:

- How to find sunsets?
- Is user annotation an issue for broadcast content holders?-
- How to pre-process entering data (830.000 hours per year) without generating too much noise?
- Challenges concern content structuring and thematic clustering

Other issues for Audiovisual Archives:

ina EXPERT
Enseignement
Formation
Recherche

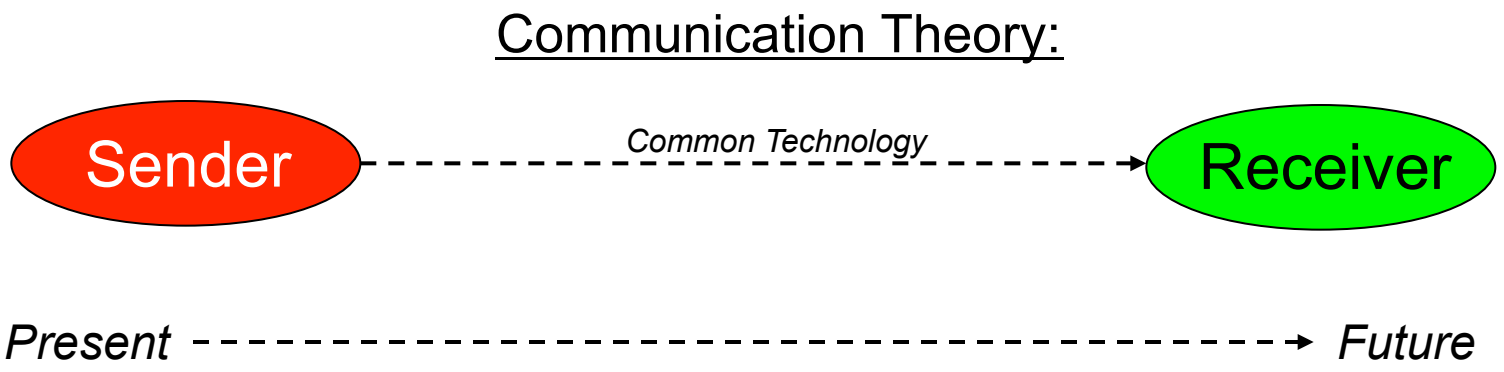
- Storage issues for the future: grids, replication, long-term migrations, archive linking for content securisation
- Ensuring the long term integrity of digital contents (security, integrity, tracability, monitoring, persistency, etc.)
- User-generated documentation and their integration in documentation schemes and practices
- Documentation and Preservation models for complex contexts
- Content Quality: does the content I'm looking at look as it initially looks?
- Perpetual expansion of storage, indexation, documentation and right management (200 000 right holders in Ina)

Migration; Multiple Definitions for the Same Word

Conservation: Keeping objects (contents) to make them available in the future

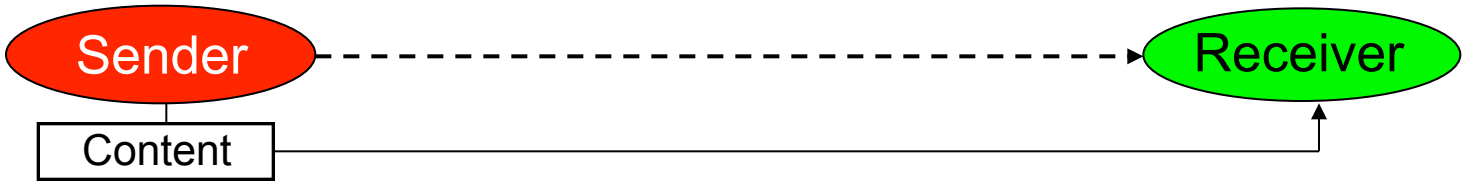
Preservation: Communication with the future

Digital Preservation: Only viable approach to long-term preservation of digital objects

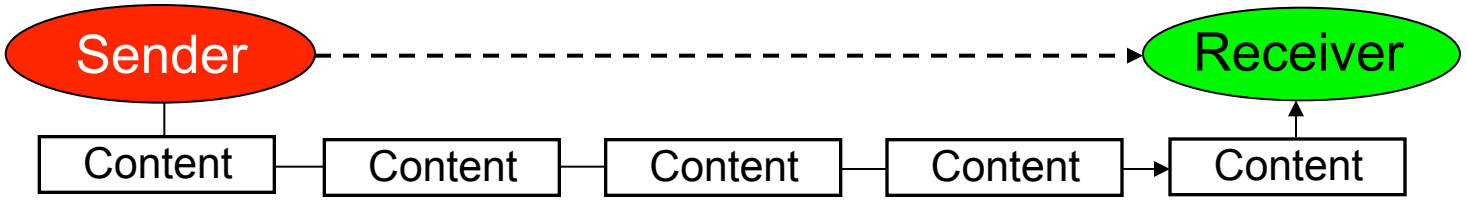


Three ways of sending Digital contents to the Future:

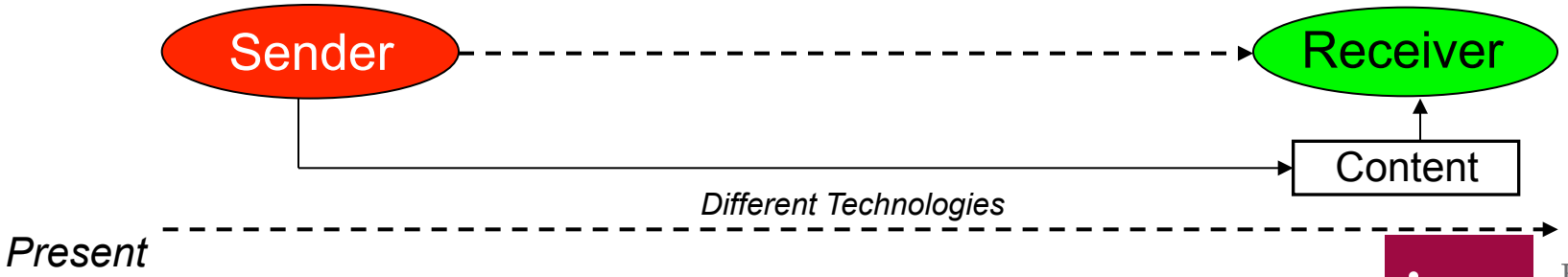
1) Change now, recover later



2) Change continuously

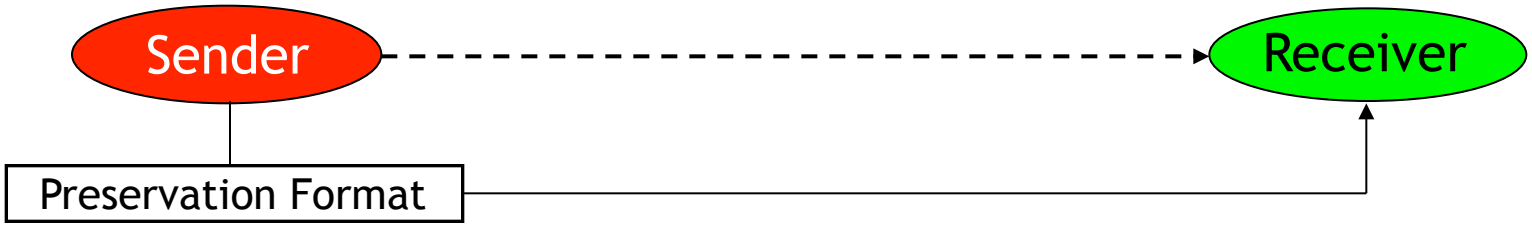


3) Don’t change, it will be solved later



Three migration approaches to Digital Content

1) Migration on ingest: change your contents to a chosen format



Implication:
Postpones migration but does not solve the problem

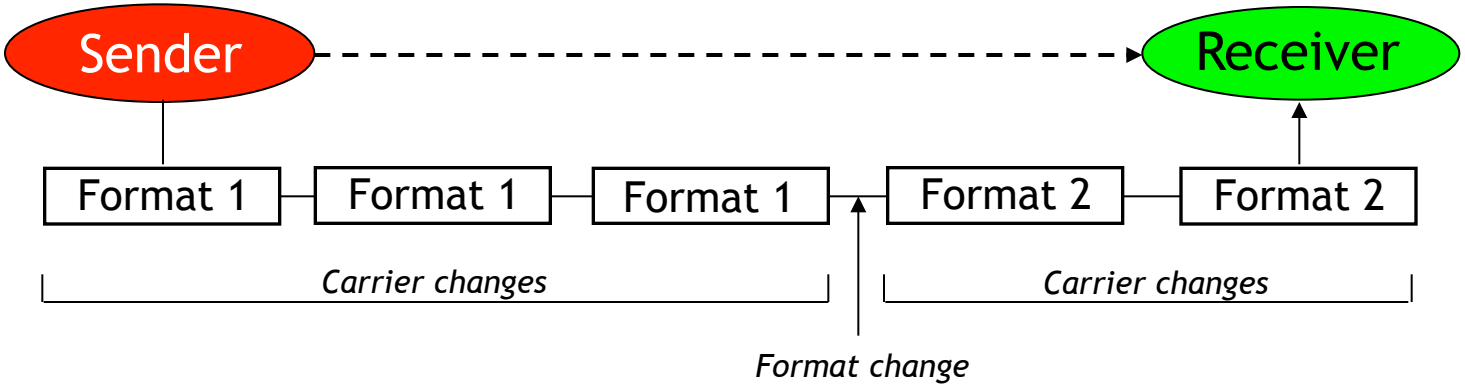
Some issues:
UPF: Universal Preservation Format
UVC: Universal Virtual Computer

Efficient solutions for repositories

Present ---------- *Different Technologies* ----------→ *Future*

Three migration approaches to Digital Content

2) Batch Migration : change when necessary (media, format or system obsolescence)



Associated concepts:

Refreshing: transfer same data to a new carrier

Integrity check: is the data the same that it is supposed to be?

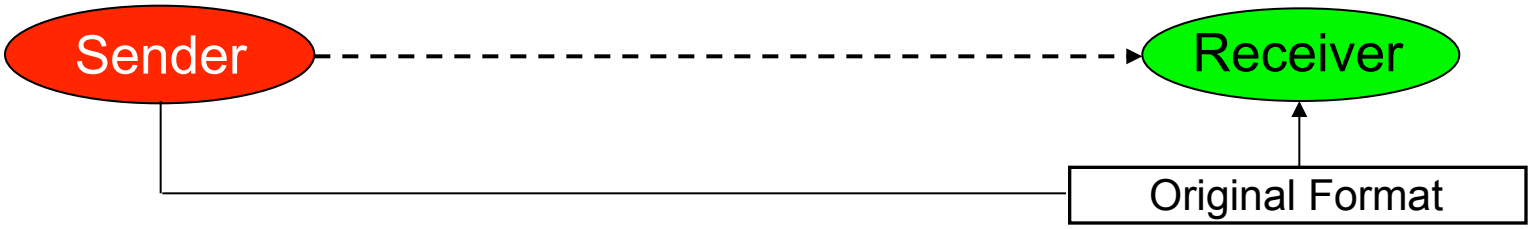
Transcoding or **Conversion**: changing formats

Efficient solution for continuously used contents!

Present ---------- *Different Technologies* ----------→ *Future*

Three migration approaches to Digital Content

3) Migration on access: postpone migration until you need it



Very useful if:

- Correct representation of information and structure as well as semantic labels on the structures present within the record (OAIS)
- Representation description of the Preservation environment

Also called:

Emulation: replicating the functionality of an obsolete system

Implication:

- Need to very precisely describe the original environment
- Need to archive format converters or develop access software

Present ---------- *Different Technologies* ----------→ *Future*

PrestoPRIME

FP7 3rd Call

Integrated Project – 42 months

Starting date : 1/1/2009- ending 30/11/2012

Partners : INA, BBC, RAI, B&G, ORF, JRS, ExLibris, Eurix, Doremi, IT Innov, Vrije Universiteit Amsterdam

Universität Innsbruck, University of Liverpool

European Digital Library Foundation

www.prestoprime.eu

SEVENTH FRAMEWORK PROGRAMME

ina EXPERT

Enseignement
Formation
Recherche

PrestoPRIME Summary

R&D for the long-term preservation of digital audiovisual objects, programmes and collections. Increasing access by integrating the media archives with European on-line digital portals in a digital preservation framework.

O1 To research and develop means of ensuring the permanence of digital audiovisual content in archives, libraries, museums and other collections.

O2 To research and develop means of ensuring the long-term future access to audiovisual content in dynamically changing contexts.

O3 To integrate, evaluate and demonstrate tools and processes for audiovisual digital permanence and access.

O4 To establish a European networked Competence Centre to gather the knowledge created by PrestoPRIME and deliver advanced digital preservation advice and services in conjunction with the European Digital Library Foundation and other projects.

SEVENTH FRAMEWORK PROGRAMME

ina EXPERT
Enseignement
Formation
Recherche

Some PrestoPRIME actions

- Models and Metadata for Audiovisual long-term preservation
- Storage strategies and rule sets for preservation
- Processing and workflows for Audiovisual migration
- Content quality appraisal and risk management
- Multivalent approaches to long-term AV media preservation
- Infrastructures for AV content storage and processing
- Metadata interoperability for access
- User-generated and contextualised metadata
- Content provenance and tracking
- Audiovisual rights modelling at European level
- Integration of Archives, Libraries and user generated content
- European Audiovisual Competence Centre

Merci

ina EXPERT

Enseignement
Formation
Recherche