A leading institution at the heart of the digital society

EURECOM
Sophia Antipolis

Le projet OligoArchive

Stockage de données numériques basé sur l'ADN

Raja Appuswamy

EURECOM

(en collaboration avec les
Archives Nationales Danoises)

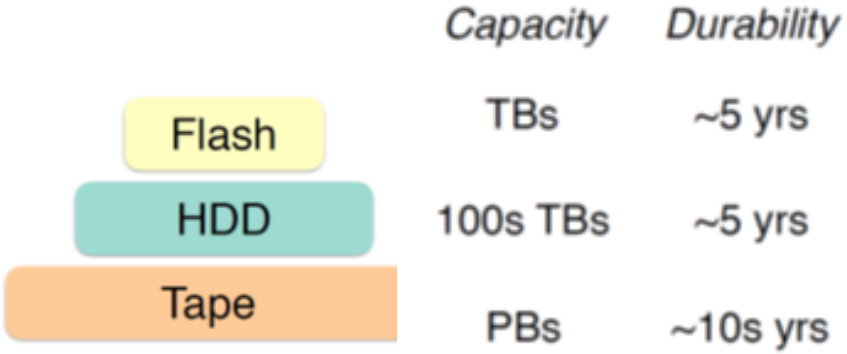
Cti
Commission des Titres d'Ingénieur

Préservation des données culturellement significatives

- **Archive Nationale Danoise**
  - Préservation des données créées et rétro-numérisées depuis 1970
- **Dessins numérisés du roi Christian IV**
  - Les dessins remontent à la période 1583-1591
  - Matériel classé comme ayant une importance nationale unique
  - Partie d'une large unité d'archivage (TIFF, métadonnées) stockée dans des bases de données

Les défis de la préservation numérique

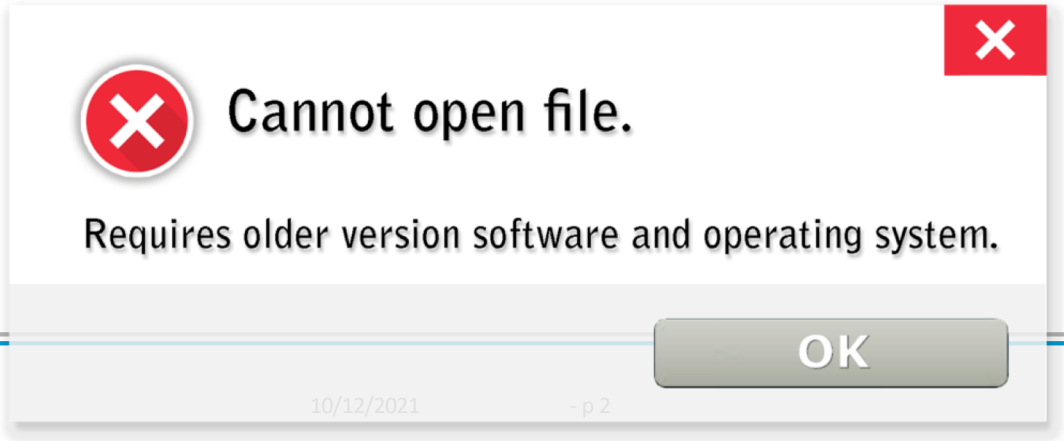
Dégradation des médias



Obsolescence des médias

	Tape Drives				
Version	LTO-6	LTO-5	LTO-4	LTO-3	LTO-2
LTO6	Read/Write				
LTO6 WORM	Read/Write				
LTO5	Read/Write	Read/Write			
LTO5 WORM	Read/Write	Read/Write			
LTO4	Read	Read/Write	Read/Write		
LTO4 WORM	Read	Read/Write	Read/Write		
LTO3		Read	Read/Write	Read/Write	
LTO3 WORM		Read	Read/Write	Read/Write	
LTO2			Read	Read/Write	Read/Write
LTO1				Read	Read/Write
Cleaning Tape	Supported	Supported	Supported	Supported	Supported

Obsolescence des formats



Préservation en cours basée sur la migration

28 Apr 2017 | 15:00 GMT

The Lost Picture Show: Hollywood Archivists Can't Outpace Obsolescence

Studios invested heavily in magnetic-tape storage for film archiving but now struggle to keep up with the technology

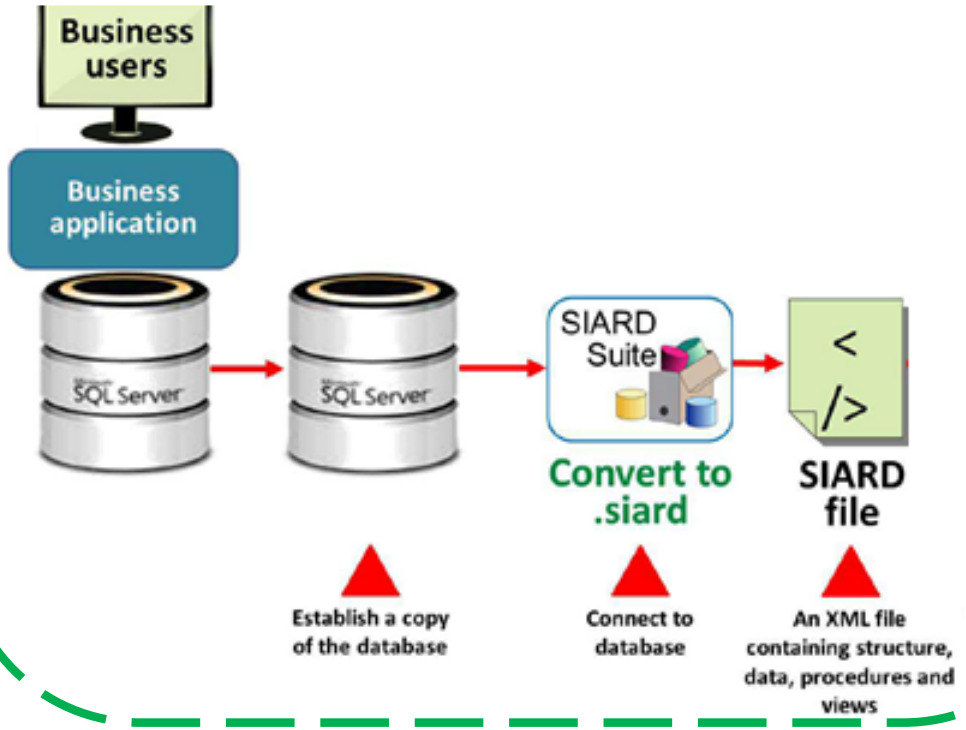
By **Marty Perlmutter**

"There's going to be a large dead period," he told me, "from the late '90s through 2020, ***where most media will be lost.***"

Vers une préservation passive et holistique

DILCISBoard/**SIARD**

SIARD (Software Independent Archiving of Relational Databases) - an open file format for the long-term archiving of relational databases

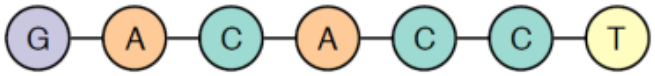


DNA molecules

Four nucleotides:

- A Adenine
- C Cytosine
- G Guanine
- T Thymine

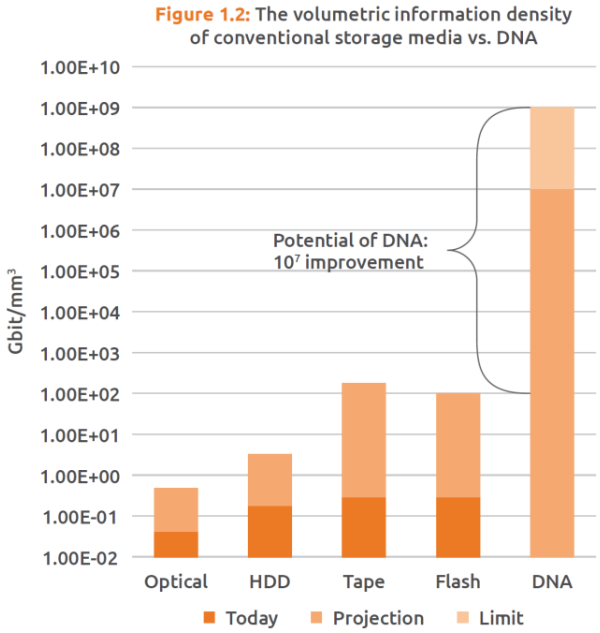
DNA strand (oligonucleotide) is a linear sequence of these nucleotides



Résoudre l'obsolescence des formats par les normes

Résoudre les problèmes médiatiques par l'ADN

Pourquoi l'ADN?



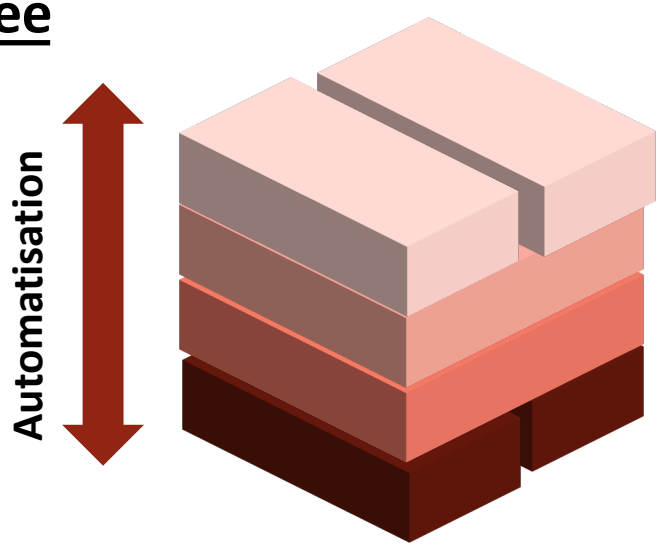
Project Oligo Archive se concentre sur l'utilisation de l'ADN comme support de stockage intelligent

OLIGO ARCHIVE

Densité 10^7 fois plus élevée


Woolly mammoth on verge of resurrection, scientists reveal

Scientist leading 'de-extinction' effort says Harvard team could create hybrid mammoth-elephant embryo in two years



Couche Application

Encodage de données structurées (base de données) et non structurées (imagerie)

Couche OS

Accès avancés (bloc, fs, …)

Couche Contrôleur

Traitement des requêtes quasi-moléculaire

Couche Media

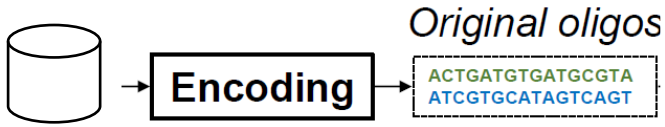
Synthèse et séquençage

Durable, éternellement pertinent

EURECOM
Sophia Antipolis

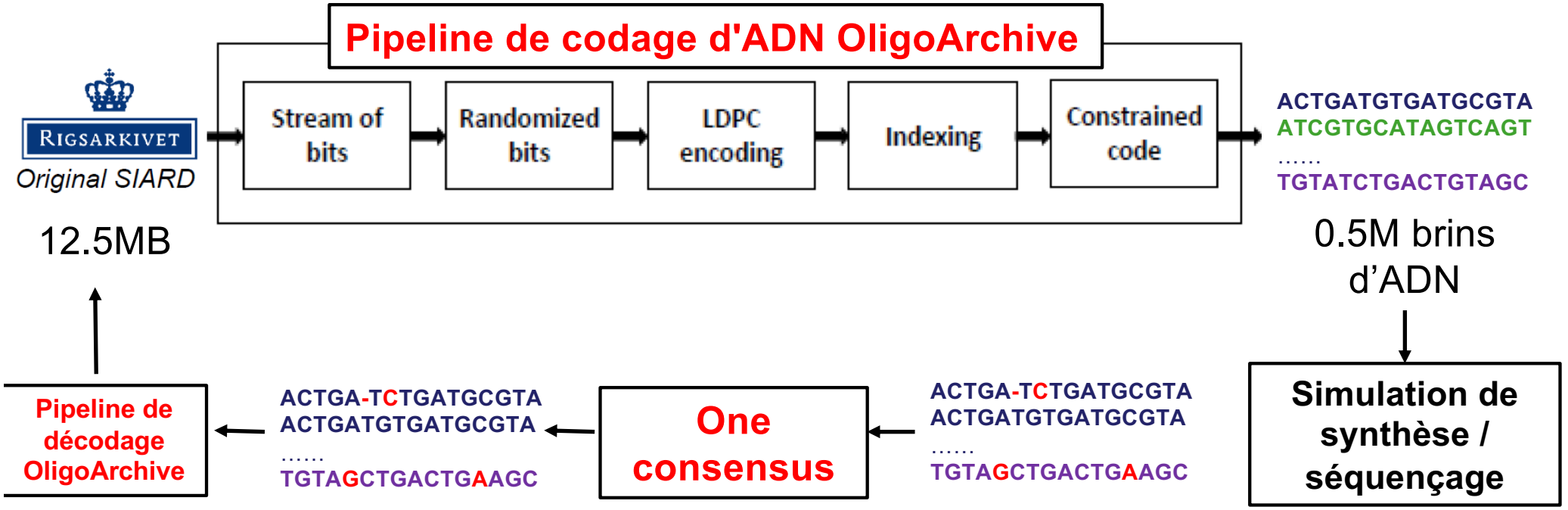
Archivage et restauration de l'ADN: Défis

- **Chaque ADN est limité à quelques centaines de nucléotides**
  - Des données réparties sur des millions d'ADN
- **Tous les ADN ne sont pas créés égaux**
  - Limitations du contenu G-C, homopolymères
- **L'ADN n'a pas d'adressage**
  - Besoin d'ajouter des informations de séquençage dans l'ADN



Erreurs biochimiques
- Substitution, insertions, suppressions
- Biais et duplication

Collaboration DNA4DNA :
ADN synthétique et les Archives Nationales Danoises



- **Troisième plus grande expérience universitaire en cours**
  - 200Mo (Microsoft/UW), 22Mo (Blawat et al.)
  - Capacité proche de la densité de stockage:1.73 bits/nt
  - 12,5 Mo en quelques nanogrammes

Avocat du diable: problèmes ouverts

- **Prix : La synthèse d'ADN est 10^7 fois plus chère que la bande magnétique (10$/To pour la bande magnétique vs 100M$/To pour l'ADN)**
  - Nouvelles techniques de synthèse en cours de recherche
- **Automatisation, performances de synthèse et séquençage**
  - La synthèse/le séquençage est laborieux et lent
  - Débit ADN O(Kb/s) par rapport aux Mo/s de la bande magnétique
- **L'ADN ne résout pas l'obsolescence des supports/formats**
  - SIARD aide à l'obsolescence des formats
    - Qu'en est-il des formats non standard ?
  - Qui préserve le décodeur (obsolescence des médias) ?
    - Collaboration en cours avec EUPALIA sur l'émulation

Conclusion

- **Défis de la préservation numérique à long terme**
  - Les bases de données souffrent de l'obsolescence des formats de fichiers
  - La bande souffre de la dégradation et de l'obsolescence des supports
- **Approche active de la préservation numérique non durable**
  - La migration continue coûte cher
- **L'ADN offre une alternative biologique**
  - Dense, durable, pertinence éternelle
- **Oligo Archive permet l'utilisation de l'ADN comme média numérique**
  - Encodage tolérant aux erreurs, pipelines entièrement automatisés
- **ADN 4 ADN : Oligo Archive + Archives Nationales Danoises**
  - Formats de fichiers standards (SIARD) pour résoudre l'obsolescence des formats
  - ADN synthétique pour résoudre la dégradation et l'obsolescence des médias

UAG
UGA
UAA