Huma-Num
la TGIR des humanités numériques

Archivage des données à Huma-Num

Michel Jacobson
TGIR Huma-Num

GRICAD & SARI : journée consacrée à l'archivage numérique des données de recherche
Grenoble - 20 novembre 2019
https://dataarchivage.sciencesconf.org/

cnrs Aix*Marseille université CAMPUS CONDORCET Paris-Aubervilliers

Plan

- Huma-Num en quelques mots
- Historique de l’activité d’archivage à HN
- Etat des lieux des projets en cours

Présentation d'HN

- Les missions d'Huma-num
  - Accompagner l'évolution des communautés SHS dans contexte du numérique et de la science ouverte
  - Outiller les programmes de recherche SHS dans la logique de libre accès aux données, de données documentés scientifiquement (FAIR)
  - Participer à la construction des infrastructures internationales (en incluant les communautés SHS nationales)

Le cadre

Label donné par le ministère de « très grande infrastructure de recherche » (TGIR)

HUMA-NUM

Humanités numériques

Huma-Num est une très grande infrastructure (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales.

Pour remplir cette mission, la TGIR Huma-Num est bâtie sur deux piliers :
- des consortiums, composés de projets et équipes de recherche financés par Huma-Num et qui partagent un intérêt commun pour des objets scientifiques ;
- un dispositif technologique unique, déployé à l'échelle nationale et fondé sur un vaste réseau de partenaires.

Cette infrastructure offre une grande variété de plateformes et d'outils pour stocker (Huma-Num-Box), traiter (Boîte à outils partagée), partager (NAKALA) et lier (ISIDORE) les données de la recherche. Huma-Num porte la participation de la France dans deux infrastructures Européennes de type ERIC (European Research Infrastructure Consortia) : DARIAH (Digital Research Infrastructure for the Arts and Humanities) et CLARIN (Common Language Resources and Technologies Infrastructure). Par ailleurs, Huma-Num est également impliquée dans d'autres projets Européens (H2020) et internationaux.

RELATIONS AVEC LES ACTEURS ÉCONOMIQUES ET/OU IMPACT SOCIO-ÉCONOMIQUE

La TGIR travaille avec le monde de l'industrie de la connaissance, du search engine et du big data dans le but d'améliorer l'appropriation, par les communautés SHS, des enjeux de l'économie de la donnée numérique.

DONNÉES

Estimation du volume de données stockées en 2017 : 1 Po

Volume de données stockées prévisible à 5 ans : 4 Po

Hors contraintes légales, l'accessibilité des tiers aux données est : complète

Coût complet
—
2,9 M€ en 2016

Personnels
—
12,2 ETPT en 2016

Dimension internationale

DARIAH, ESFRI Landmark

Directeurs : Laurent Romary, Frank Fischer, Jennifer Edmond

Pays coordinateur : France

Pays partenaires : DE, AT, BE, CY, HR, DK, GR, IE, IT, LU, MT, NL, PL, PT, SI, RS, SE

Site internet : www.dariah.eu

CLARIN, ESFRI Landmark

Directrice : Franciska de Jong

Pays coordinateur : Pays-Bas

Pays Partenaires : AT, BG, CZ, DK, EE, FI, DE, GR, IT, LT, NO, PL, PT, SI, SE, UK, FR (observateur)

Site internet : www.clarin.eu

Catégorie : TGIR

Type d'infrastructure : Distribuée

Localisation : Paris

Localisation des autres sites : Villeurbanne

Établissement français porteur : CNRS

Directeur de l'infrastructure en France : Olivier Baude

Création :	Exploitation :
2013	2013

Tutelles / Partenaires : AMU, Campus Condorcet

Contact en France :
direction@huma-num.fr

www.huma-num.fr

Présentation d’HN

- La TGIR Huma-Num regroupe des ressources humaines et technologiques à l’échelle nationale et européenne en s’appuyant sur un important réseau de partenaires et d’opérateurs.
  - Une équipe de ~18 personnes
  - Des consortiums (regroupement d’acteurs des communautés scientifiques)
  - Un dispositif technologique permettant le traitement, la conservation, l'accès et l'interopérabilité des données de la recherche. Ouvert à l'ensemble des programmes de recherche de l'enseignement supérieur et de la recherche.
  - Des partenaires et opérateurs tels que le réseau des MSH ou le CINES
- La TGIR Huma-Num porte la participation de la France dans les ERIC (European Research Infrastructure Consortium) DARIAH et CLARIN en coordonnant les contributions nationales. Elle est également impliquée dans cinq projets H2020 : Parthenos, Humanities at Scale (terminé), TRIPLE, SSHOC et EOSC-PILLAR.

Les communautés

CAHIER
(Corpus d'Auteurs pour les Humanités : Informatisation, Édition, Recherche)

CONSORTIUM ARCHIVES DES ETHNOLOGUES

CORLI
(Corpus, langues, interactions)

3D-SHS

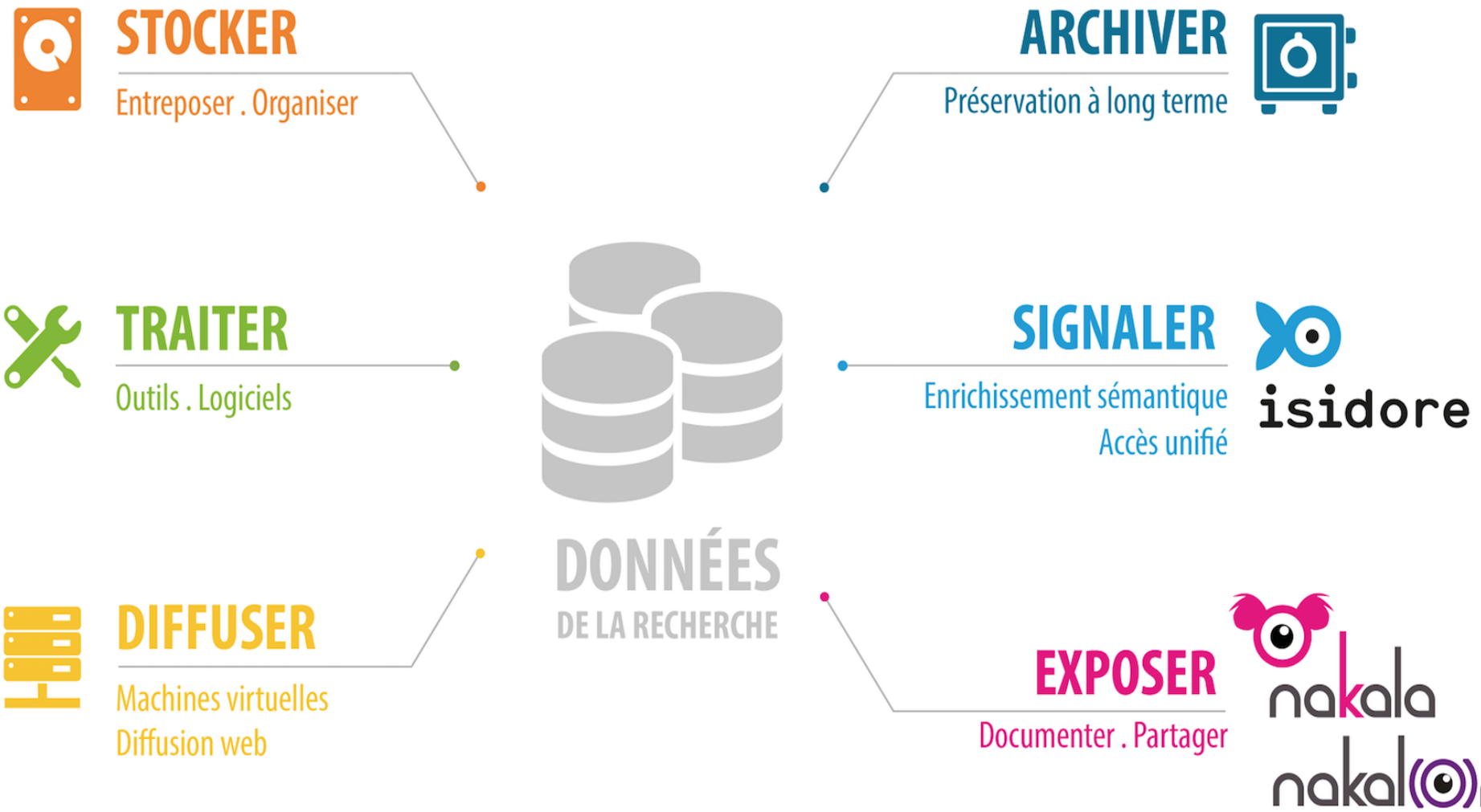
Paris Time Machine
Données géo-historiques et carto-historiques

COSME
(Consortium sources médievales)

IMAGEO
(Cartes et photographies pour les géographes)

MASA
(Mémoires des archéologues et des sites archéologiques)

Les services



L’offre de préservation à long terme à HN

ARCHIVER

- Huma-Num accompagne les projets de préservation à long terme
- Liens entre les producteurs de données et le CINES
- Suggestion de nouveaux formats et prise en compte (e.g. formats pour la 3D)
- Liens avec la communautés des archivistes

CINES
Centre Informatique National de l'Enseignement Supérieur

HN Huma-Num

Historique du projet archivage

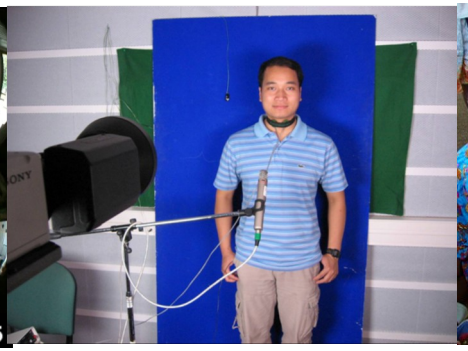
- Programme archivage du TGE-Adonis (2008-2010)
  - Pré-étude du CERN pour savoir sur quel service adosser ce programme
  - Choix de 2 centres de calcul (CINES et CC-IN2P3)
- Équipe du projet pilote associant :
  - TGE-Adonis (maîtrise d'ouvrage)
  - En maîtrise d’oeuvre :
    - CINES (opérateur d'archivage)
    - CC-IN2P3 (site secondaire + fonctionnalités d'accès)
  - DAF (tutelle ministérielle)
  - CRDO (pour tester un type de données)
  - Un consultant du CNES spécialiste de la norme OAIS

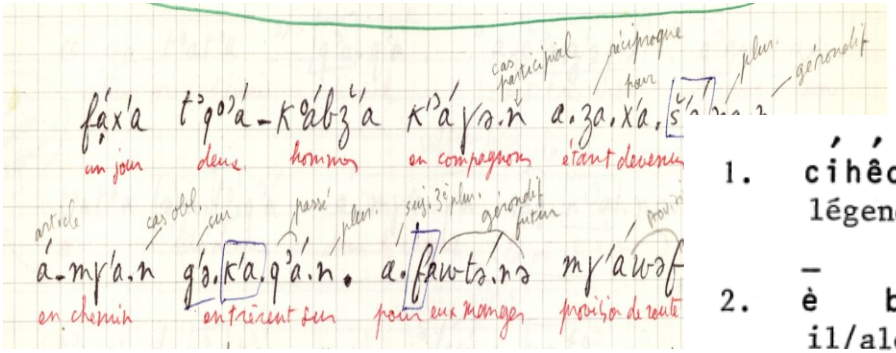
Historique du projet archivage

Raisons du choix du pilote

- Une communauté déjà organisée
  - des laboratoires de linguistiques regroupés en fédérations
  - Un centre de ressources numérique pour la description de l'oral CRDO (2 antennes : Paris → COCOON ; Aix → SLDR → Equipex Ortolang)
- Un modèle de métadonnées : vocabulaires OLAC (Open Language Archives Community)
- Une réflexion sur les pratiques : « Corpus Oraux : guide des bonnes pratiques »
- Une typologie étendue de ressources
  - Enregistrements (audio + vidéo)
  - Annotations (XML, PDF, textes, images)

Enregistrements/Annotations





1. cíhêdéé ka cíhê bwaaoléé pwö- a péi kúcúkúcú
légende/et/parler/aigle/dessus-/*n.*le/rocher/Kucukucu/

2. è bwö mú ka è bwö cìnì wíi ìhîm
il/alors/demeurer/et/il/alors/griller/manger/bancoulier/

\ref T120-C12 003

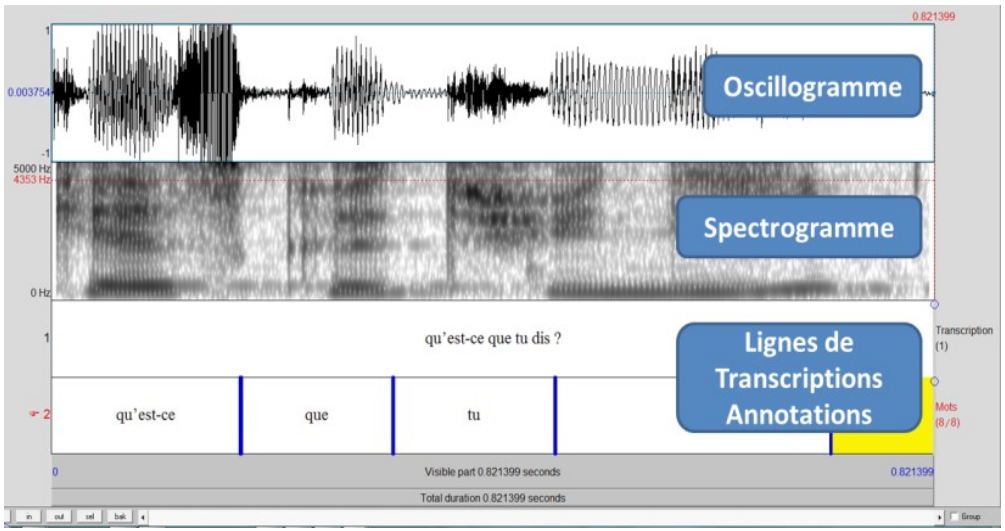
\tx	**mè**	**né**	**hǫ́gǫ́,**	**tɔ̀rìɔ̀ŋ**	**ʔą̀**	
\rm	mɔ̀	né	hǫ́gǫ́	tɔ̀rìɔ̀ŋ	ʔą̀	
\ma	chose	être	comme ça	petit grillon	3S	
\am	N	PRED	ADV	N	PR	

\tx	**gɔ̀nà**			**zǫ́**	**kǫ́ą̀,**		
\rm	BHa-	gɔn	-H	zǫ́	kǫ́	ʔą̀,	
\ma	Acc-	découper	-D	herbes	de	3S	
\am	MV-	V	-FCT	N	FCT	PR	

\tx	**gɔ̀nà**			**zǫ́**	**gásá**	**zǫ́**	
\rm	BHa-	gɔn	-H	zǫ́	gásá	zǫ́	
\ma	Acc-	découper	-D	herbes	grand	herbes	
\am	MV-	V	-FCT	N	AV	N	

\tx	**hé**	**mè**	**né**	**gbàmbɔ̀ndɔ́**	**mèí**		**gá.**
\rm	hé	mɔ̀	né	gbàmbɔ̀ndɔ́	mè	-íí	gá.
\ma	comme	chose	être	herbes sp	là-bas	-anaph	comme
\am	SUB	N	PRED	NPR	ADV	-MOD	SUB

\tr C'est ainsi que le Grillon il a délimité son territoire de chasse, il a délimité un territoire, un grand territoire, comme qui dirait Gbambondo là-bas.



Historique du projet archivage

Les premiers travaux

- Faire monter en compétence l'ensemble des acteurs sur le modèle OAIS
- Étudier les formats utilisés par la communauté choisie (l’oral)
- Établir un guide méthodologique pour le choix de formats numériques pérennes dans le contexte de données orales et visuelles généralisable à d’autres contextes.

Historique du projet archivage

Étude sur les formats. Quelques particularité des formats audio-visuel

- Encodage vs formatage : MP3
- Hiérarchies de dépendances : WAV > BWF
- Formats conteneurs : MP4

Des critères pour évaluer les formats

- Ouvert
- Normalisé
- Largement utilisé
- Existence d'outils de contrôle de la conformité du format avec sa spécification

Une méthodologie réutilisée pour les études suivantes (par ex. les formats PDF)

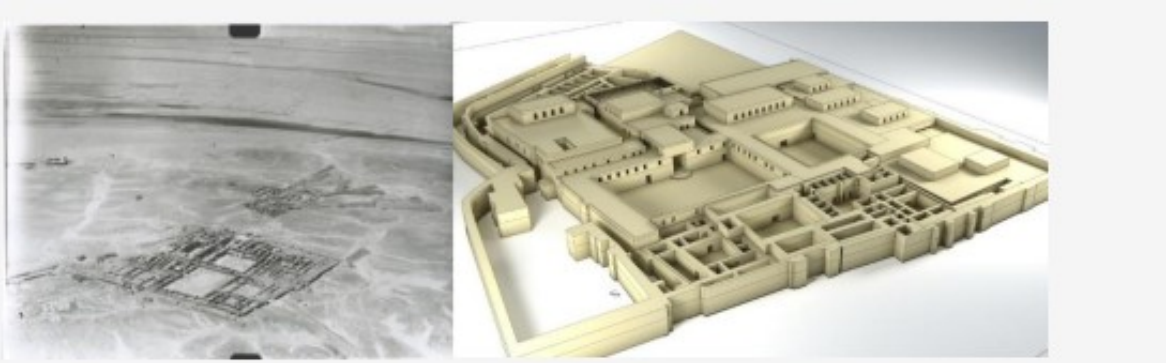
Historique du projet archivage

Les premiers travaux (suite)

- Spécifications de nouvelles fonctionnalités à ajouter dans la plate-forme d'archivage du CINES (PAC)
  - Nouvelles transactions : mise à jour des métadonnées, versions
  - Attribution d'identifiants pérennes indépendants de la plate-forme : choix de ARK
- Spécifications de fonctionnalités d'accès sur le CC-IN2P3
  - Choix de Fedora : Piste abandonnée
  - Les Centres de ressources numériques conservent les fonctionnalités d'accès
- Ajouts d’informations de gestion
  - Communicabilité
  - Durée de conservation et sort final

Les autres projets d’archivage

Ouverture à d’autres projets et types de ressources : Les données 3D d’Archéovision

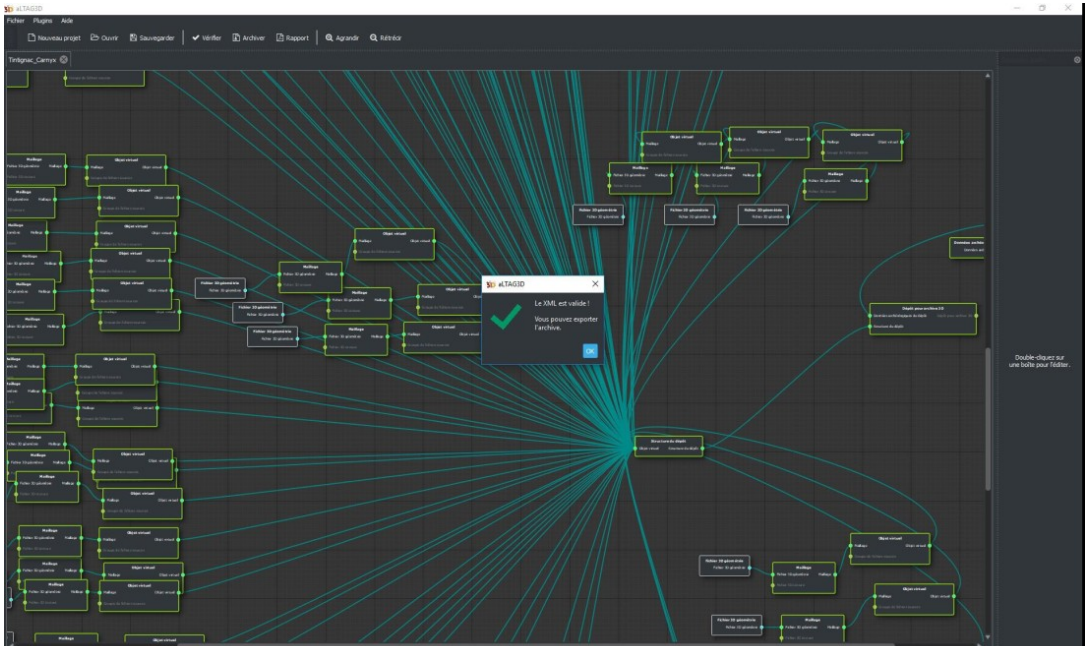


Restitution du palais de Mari

Buste d’Akhenaton

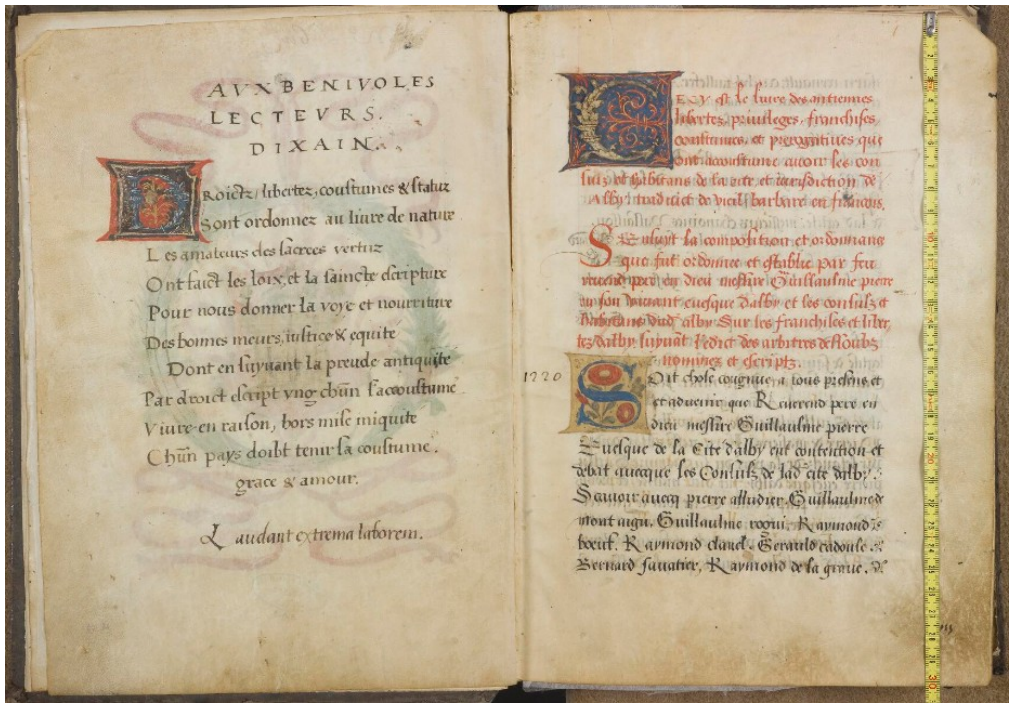
Les données 3D d’Archéovision

- Choix des formats de représentation des données (collada, ply)
- Travaux du consortium 3D
  - Conception d’un schéma de métadonnées métier (MDACST3D) pour les données 3D archéologiques du patrimoine culturel
  - Développement d’un logiciel d’archivage (aLTAG3D) de préparation des paquets à archiver au CINES



Autres fonds

- Les manuscrits anciens de l'IRHT



- Un million de fichiers TIF occupant de l'ordre de 50 To.
- Conversion en JPEG2000
- Récupération des métadonnées techniques encapsulées dans les fichiers image pour les exprimer en XML/RDF
- Récupération des métadonnées documentaire et scientifiques formatées en XML/TEI
- Description de l'organisation des images entre elles pour former des manuscrits en XML/METS

Autres fonds

- Les fonds de la photographique de l’École Française d'Extrême-Orient (EFEO)
  - Plus de 200 000 clichés pris lors des fouilles et des missions organisées par l'École depuis sa création en 1900 (Cambodge, Vietnam, Laos…)
  - Premiers versements en 2009 reprise en 2015 pour reverser après avoir enrichi les métadonnées et développé un outil de versement
- Les fonds d’enregistrements de la médiathèque de la MMSH
- Les enquêtes quantitatives et qualitatives du Centre de Données Socio-Politiques (CDSP de SciencesPo)
  - Formats divers : csv, pdf, wave, XML (EAD, TEI)…
  - Utilisation du protocole SEDA pour ses échanges avec le CINES
- Les éditions de la chaîne OpenEdition Books (en cours...)