Présentation - services - stockages

Joël Marchand

TGIR Huma-Num - CNRS UMS 3598

6 octobre 2016

Huma-Num

Présentation

Faciliter le « tournant numérique » de la recherche en sciences humaines et sociales dans la production et la réutilisation de données numériques.

Développer l'appropriation par les communautés scientifiques du cycle de vie des données numériques.

Proposer des services pour les données au juste niveau et au bon moment.


CONSORTIUMS
Concertation
Bonnes pratiques
Expertise
PROJETS DE RECHERCHE
OUTILS ET SERVICES
Stocker . Traiter . Diffuser
Exposer . Signaler . Archiver
RESEAU INTERNATIONAL
Partage . Expertise

Huma-Num

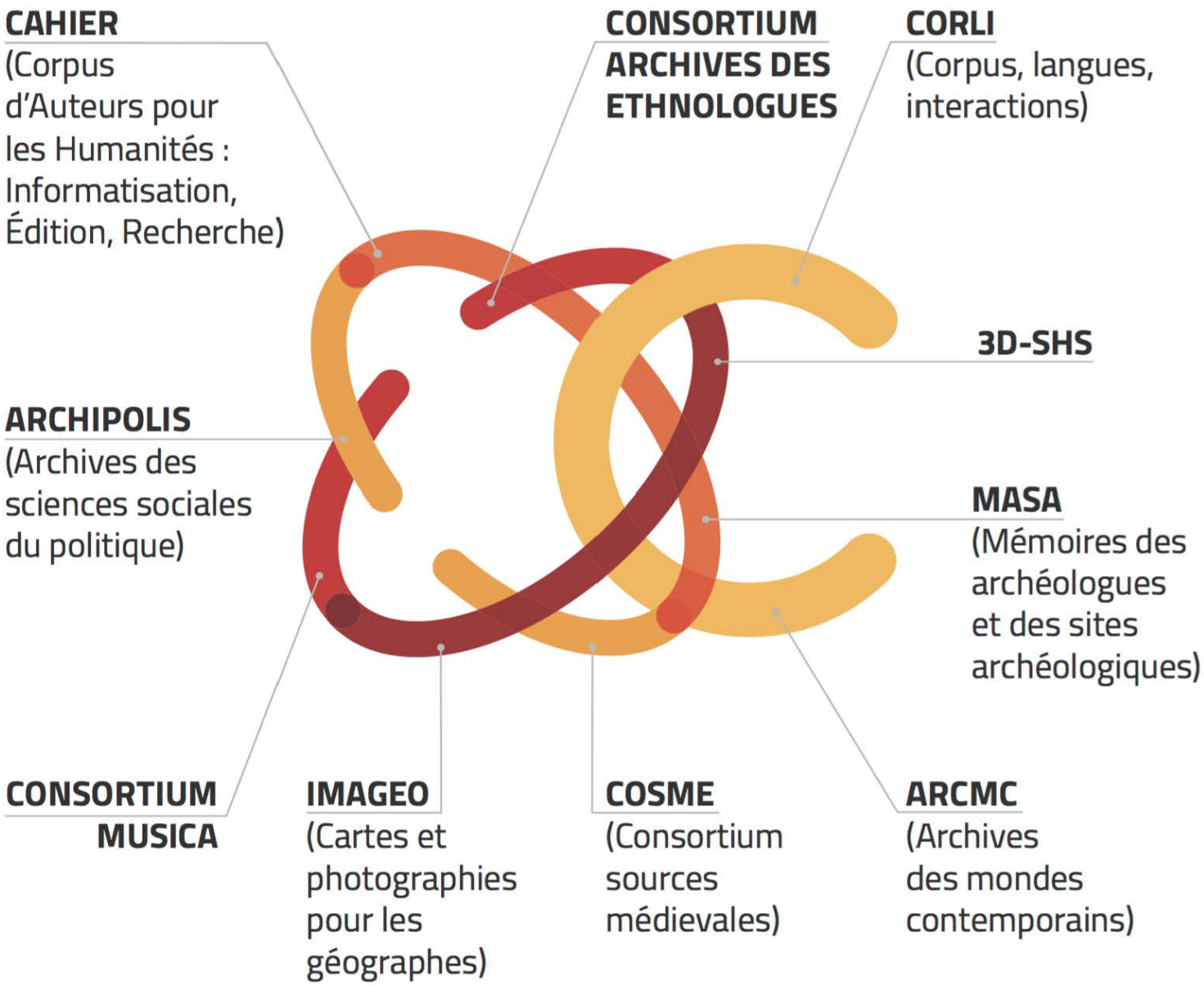
Principes de fonctionnement

- Démarche bottom-up
- De nombreux enseignants-chercheurs sont impliqués
- Ils demandent des outils, des méthodes, des normes
- Avec les consortiums - qui organisent - nous créons des services dédiés et nous les exploitons

Huma-Num

Consortiums

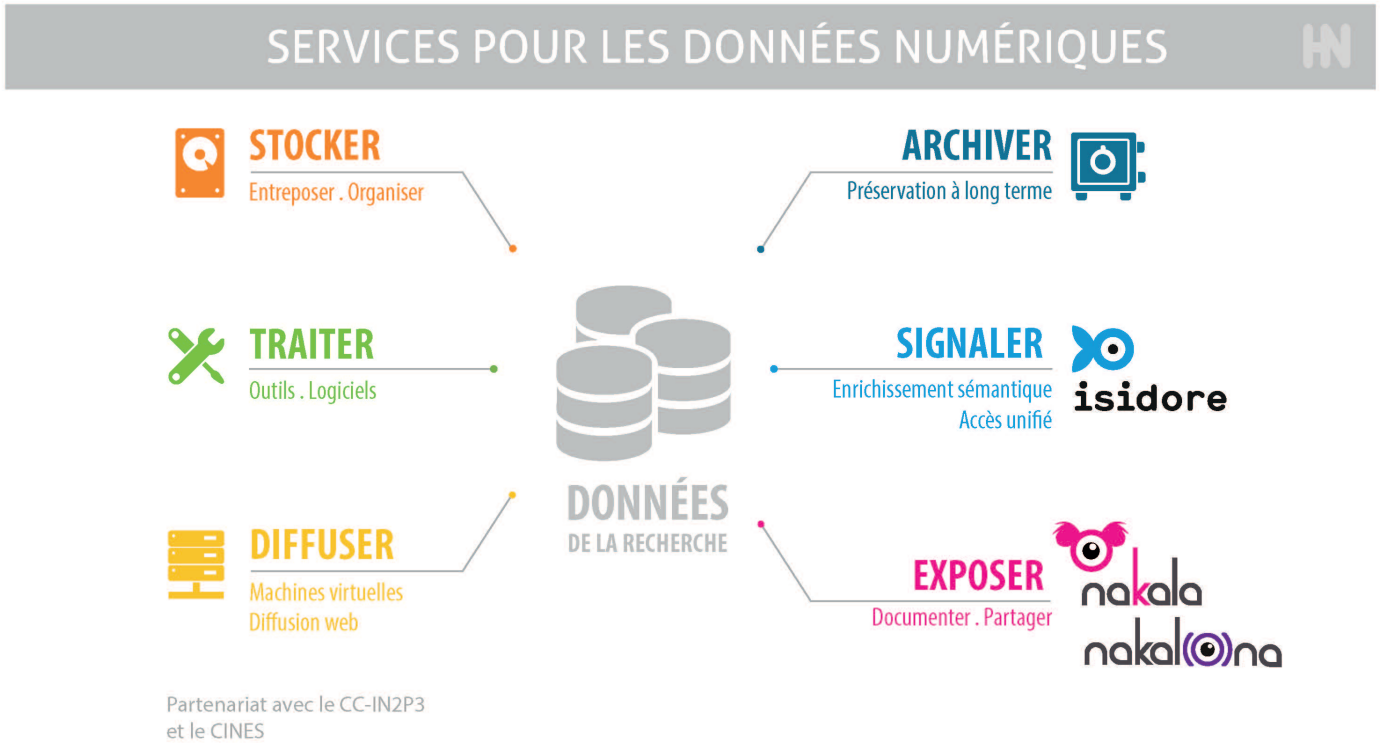
- Regroupement d'unités et équipes de recherche autour de thématiques et d'objets communs
- 10 consortiums labellisés
- 120 équipes de recherche : UMR, EA
- Un réseau : + 300 chercheurs, ingénieurs, etc
- + 50 actions publiques : formations, ateliers, séminaires
- Ancrage dans les Maisons de Sciences de l'Homme (MSH)
- Soutien financier de la part d'Huma-Num


CAHIER
(Corpus d'Auteurs pour les Humanités : Informatisation, Édition, Recherche)
CONSORTIUM ARCHIVES DES ETHNOLOGUES
CORLI
(Corpus, langues, interactions)
3D-SHS
ARCHIPOLIS
(Archives des sciences sociales du politique)
MASA
(Mémoires des archéologues et des sites archéologiques)
CONSORTIUM MUSICA
IMAGEO
(Cartes et photographies pour les géographes)
COSME
(Consortium sources médievales)
ARCMC
(Archives des mondes contemporains)

Structure et activités internationales

- TGIR = Très Grande Infrastructure de Recherche
- Comité de pilotage + conseil scientifique
- Paris et Villeurbanne
- Budget : 1,4 M€
- 11 personnes
- La TGIR porte la participation française dans deux ERICs et est impliquée dans deux projets H2020

Services



Solutions pour entreposer et organiser les données numériques dans un espace sécurisé professionnel. En partenariat avec le CC-In2p3.	Solutions adaptées aux besoins de transformations ou d'analyses sur les données : logiciels de traitement, de visualisation, d'encodage, SIG, puissance de calcul, etc.	Différentes solutions pour diffuser les données (packs logiciels et machines virtuelles).	Conseil et accompagnement dans le processus d'archivage à long terme : aide aux choix des formats de données, normalisation des métadonnées, transfert dans le système d'archivage du CINES.	Service multilingue de référence pour le signalement, l'enrichissement, et l'accès aux données en libre accès de la recherche en sciences humaines et sociales.	Entrepôt interopérable et sécurisé pour déposer et partager tout type de données. Outil d'exposition.

Huma-Num

Infrastructure

- Hébergement au CC-IN2P3 : salle, réseau, sauvegarde, iRods
- 2 baies
- Infrastructure en propre
- 20 serveurs (1U - R620/R630 - 256 Go de RAM - Linux)
- 100 machines virtuelles (KVM libvirt/virsh/virt-manager)
- 1 NAS NetApp 2554 (+ 65 To nets utiles)
- 1 firewall PaloAlto 4020 (niveau 7 avec détection signatures d'attaque)
- Administration en interne (3 ETP)

Huma-Num

Traiter - 1/2

- Mise à disposition de toutes les applications libres demandées
  - Langages de programmation et scripts : C, C++, PHP, Python, Java
  - Moteurs de bases de données SQL : MySQL, PostgreSQL, PostGIS
  - Logiciels XML et autres bases de données : BaseX, eXist, couchdb, mongodb, redis, serveur FileMaker
  - Serveurs d'applications Java : Tomcat, Jetty
  - Serveurs de Triplestores RDF : Virtuoso, Sesame
  - Moteurs de recherche : Elasticsearch, Solr
  - Outils de SIG : QGIS, Geoserver
  - Calcul statistique : R

Traiter - 2/2

- Mise à disposition de certains logiciels commerciaux
- En mode jeton (utilisation sur le poste client)
  - oXygen (éditeur XML)
  - ESRI ArcGIS for desktop
  - SAFE FME Desktop (outil ETL)
  - Autodesk Infrastructure Design (conception 3D)
- En mode serveur (exécution sur notre infrastructure)
  - Kakadu (conversion JPEG2000)
  - Sorenson Squeeze Server (conversion de videos)
  - Dynmap (serveur WebMapping)
  - ESRI ArcGIS Pro et ArcGIS Online (cartographie, analyse et diffusion de cartes)
- Méthodes d'accès dépendant de chaque application : interface Web, client lourd, accès SSH, etc

Huma-Num

Exposer - 1/2

- Nakala : service conçu et porté par la TGIR
- Entrepôt sécurisé de données numériques
- Accessibilité directe aux données
- Citabilité garantie dans le temps
- Fonctionnalités
  - Identifiants pérennes (handle)
  - Préparation à l'archivage (CINES)
  - Vocabulaire des métadonnées : Dublin Core (DCTerms)
  - Interopérabilité OAI-PMH
  - Stockage natif en RDF / triple-store (réservoir)

Huma-Num

Exposer - 2/2

- Usages
  - Outil de chargement par lots
  - API pour les gestionnaires de données (ajout/modification)
  - API en Sparql pour les webmasters
  - Fédération d'identités
- 60 projets - 640 Go de données - 55 000 fichiers

Huma-Num

Diffuser - 1/3

- Cluster Web mutualisé Apache/PHP/Python/Java
- Moteurs de BDD SQL : MySQL/PostgreSQL
- Sites essentiellement en PHP/MySQL
- Beaucoup de CMS : Omeka (50), Drupal (39), Wordpress (29), etc
- Mais aussi : Java/Tomcat (35), BaseX (10), etc
- Des développements maison
- + 200 sites

Huma-Num

Diffuser - 2/3

- Quand besoin de plus d'outils et plus d'autonomie : machines virtuelles
- Fourniture de VM à façon en mode IAAS ou PAAS
- 30 VM à ce jour (divers Linux, Windows)

Huma-Num

Diffuser - 3/3

- Souhait fort : dissocier la conservation de la donnée et de son contexte, des outils qui la diffusent
- Couplage d'outils de diffusion avec l'entrepôt Nakala : service Nakalona avec le CMS Omeka (gestion de bibliothèques numériques) en mode SAAS

Huma-Num

Archiver

- Notre rôle : intermédiaire entre projets de recherche et CINES
- Accompagnement des projets pour maturation des données, préparation à l'archivage et dépôt selon procédure et normes du CINES
- Financement de ces dépôts auprès du CINES

Huma-Num

Signaler - 1/2

- Premier service phare de la TGIR : moteur de recherche Isidore
- Moteur sémantique travaillant sur les méta-données
- Moissonnage par OAI-PMH de ces méta-données et des données
- Alignement et enrichissement de ces méta-données par rapport à des référentiels métier (disciplines, localisations, concepts)
- Généralistes : Rameau (fr), LCSH (en), BNE (es), Géonames
- Thématiques : Pactols, Gemet, GeoEthno

Signaler - 2/2

- Recherche par facettes
- Renvoi vers les URL des données sur leur lieu d'origine
- 3 types d'accès : Web (fixe et mobile), API, Sparql (web sémantique)
- 3 langues : français, anglais, espagnol
- + 100 producteurs - + 3 900 entrepôts - + 4,6 millions de ressources
- Ex : revues.org, Persée, Cairn, Erudit, HAL-SHS, Calames, Gallica
- Basé sur les outils AIF et AFS de la société Antidot

Stockages

Sites Web

- Alimentation des sites Web par interface SFTP
- Données finalisées
- Souvent (trop) liées à l'outil de publication (CMS)
- Très souvent pas de métadonnées
- Pas d'accès direct à la donnée
- 2.8 To de données - 20 millions de fichiers

Huma-Num

Sharedocs - 1/2

- Logiciel commercial FileRun
- Gestionnaire de fichiers entièrement Web
- Connexion possible en WebDAV
- Fonctions : partages, prévisualisation, URL de diffusion, URL courtes, étiquettes, etc
- Watch-folders : couplage avec logiciels de traitement en batch
- Nos usages : Abbyy (OCR), Sorenson Squeeze Server (conversion de videos)

Huma-Num

Sharedocs - 2/2

- Simple d'emploi - appropriation facile
- "Bureau de travail" en ligne
- Utile pour élaboration du jeu de données avant publication
- 840 utilisateurs - 122 laboratoires/équipes/projets
- 16 To de données - 4 millions de fichiers

Huma-Num

Stockage capacitif de données tièdes ou froides

- Cible
  - gros volumes : plusieurs dizaines, voire centaines de To
  - données quasi finalisées : seront peu modifiées
  - données à valeur importante : méritent une sécurité élevée
  - données constituées plutôt de gros fichiers : images, sons, videos

Huma-Num

iRods et nouvelles envies

- Jusqu'à présent : iRods
  - Usage du service iRods du CC-IN2P3
  - Service performant et économique pour nous
  - Pas de clients adaptés aux SHS (Windows, MacOS)
  - Pas de montage depuis un serveur Linux
  - Mono-présence géographique
  - 130 To de données à ce jour
- Envies
  - Faciliter l'usage : comme un NAS
  - Amener des interfaces plus faciles pour les utilisateurs
  - Politiques de sécurisation souples : historisation, durée de rétention, réplication (multi-instances), multi-supports (disque, bande)
  - Disposer de fonctions supplémentaires : présentation via interface Web, fonction d'archivage (format TAR) avec contrôle d'intégrité

Huma-Num

Stockage distribué sécurisé - 1/2

- Projet en cours : stockage distribué sécurisé
  - Une "tête" NAS dans les MSH : accès facile via le LAN
  - Fonction ’Connecter un lecteur réseau’ : comme disque externe
  - Réplication possible au fil de l’eau sur un autre lieu
  - Droits d’accès et politique de sécurité par partages (jeux de données)
  - Expression évolutive dans le temps du niveau de sécurité sur les données
  - Accès depuis n’importe quel point de présence
  - Perspective d’un réseau de données national
  - Extensibilité illimitée

Huma-Num

Stockage distribué sécurisé - 2/2

- Projet en cours : stockage distribué sécurisé "Huma-Num Box"
  - Matériels : serveurs et baies SAS Dell
  - Logiciels : Active-Circle (commercial), annuaire LDAP, réseau VPN
  - 7 serveurs sur 5 MSH et nos 2 lieux : Paris et Lyon
  - Volumétrie nette utile : 300 To
  - Gestion déconcentrée des comptes et des groupes : FusionDirectory (frontal Web sur OpenLDAP)
  - Réseau VPN : OpenVPN auj. - L3VPN RENATER demain
  - 100 To de données à ce jour

Huma-Num

Références

- www.huma-num.fr
- www.rechercheisidore.fr
- www.nakala.fr
- www.nakalona.fr
- humanum.hypotheses.org
- Twitter : @huma_num
- Demande d'informations : contact@huma-num.fr
- Demande de service : cogrid@huma-num.fr

Huma-Num