Présentation des services d'Huma-Num

Joël Marchand

TGIR Huma-Num - CNRS UMS 3598

21 mai 2019

Présentation

Missions et activités

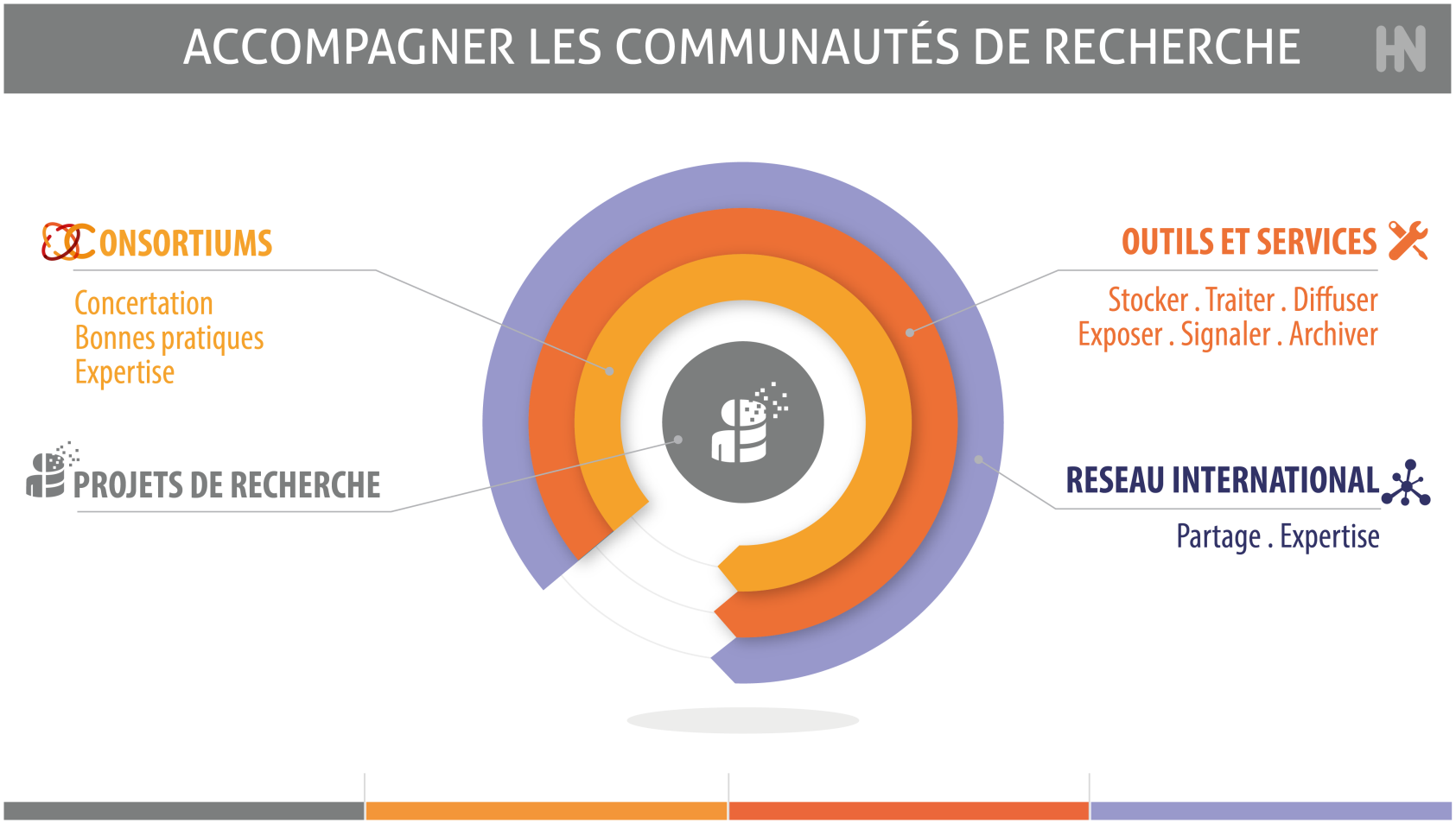
Faciliter le « tournant numérique » de la recherche en sciences humaines et sociales dans la production et la réutilisation de données numériques.

Missions et activités

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.

Missions et activités

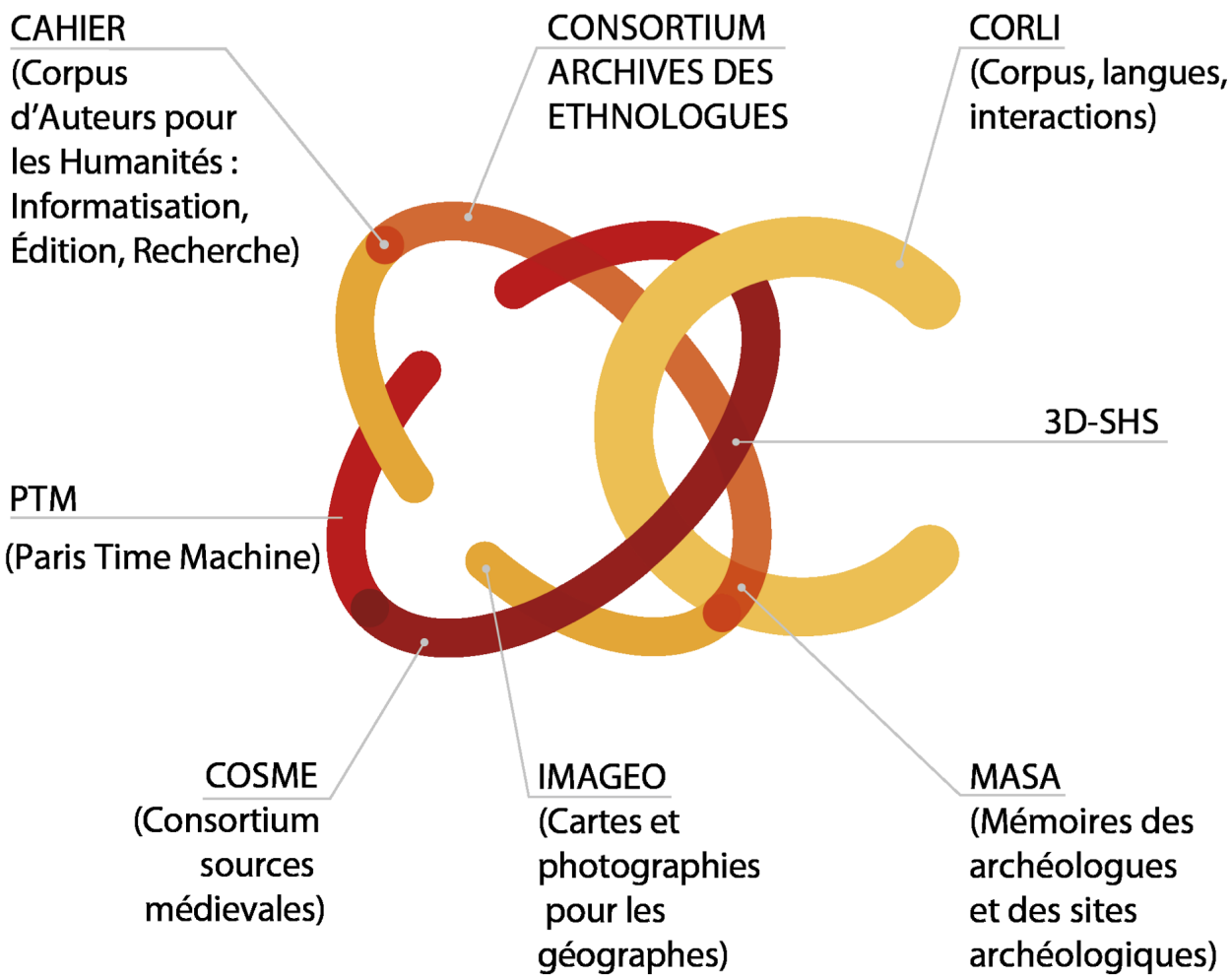


Structure et activités internationales

- TGIR = Très Grande Infrastructure de Recherche
- Comité de pilotage + conseil scientifique
- Paris et Villeurbanne
- Budget : 1,4 M€+ salaires
- 18 personnes, dont 9 BAP E
- La TGIR porte la participation française dans deux ERIC
- La TGIR est impliquée dans de nombreux projets H2020

Consortiums

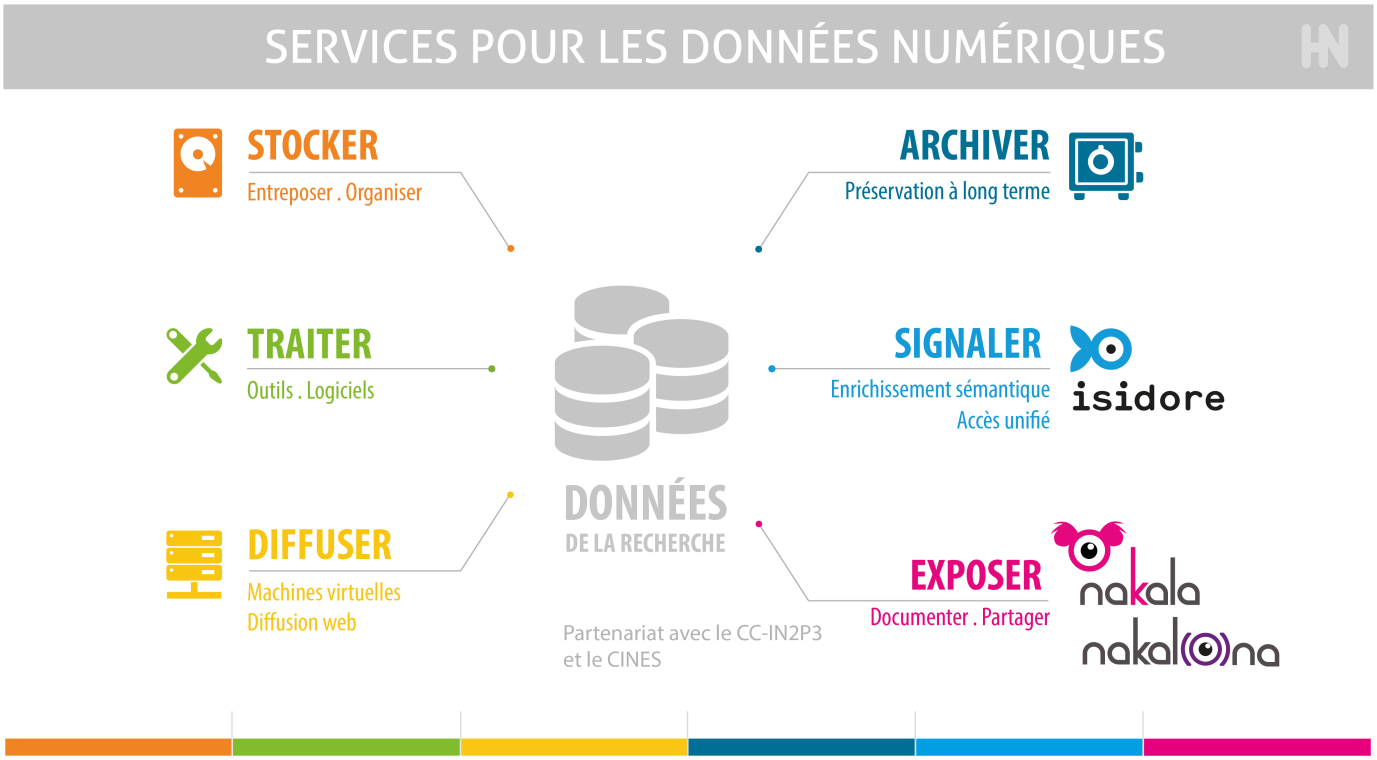
- Regroupement d'unités et équipes de recherche autour de thématiques et d'objets communs
- 8 consortiums labellisés
- 120 équipes de recherche : UMR, EA
- Un réseau : +300 chercheurs et ingénieurs
- +50 actions publiques : formations, ateliers, séminaires
- Ancrage dans les Maisons de Sciences de l'Homme (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
PTM
(Paris Time Machine)
COSME
(Consortium sources médievales)
IMAGEO
(Cartes et photographies pour les géographes)
MASA
(Mémoires des archéologues et des sites archéologiques)

Principes de fonctionnement

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

Services


SERVICES POUR LES DONNÉES NUMÉRIQUES
STOCKER
Entreposer . Organiser
ARCHIVER
Préservation à long terme
TRAITER
Outils . Logiciels
SIGNALER
Enrichissement sémantique
Accès unifié
isidore
DONNÉES
DE LA RECHERCHE
DIFFUSER
Machines virtuelles
Diffusion web
EXPOSER
Documenter . Partager
nakala
nakala
Partenariat avec le CC-IN2P3
et le CINES

Modalités d'accès

- Demandes émises librement par les porteurs de projets scientifiques
- Traitées par l'équipe opérationnelle de la TGIR
- Evaluation de l'éligibilité : appartenance à la communauté académique SHS et dans le cadre d'un programme de recherche collectif
- Mise à disposition en lien avec les interlocuteurs techniques
- Aucune contribution financière demandée
- Engagement pour aller vers une ouverture et une interopérabilité des données et/ou des métadonnées

Quand venir nous voir ?

- Avant le dépôt de demande ANR ou Europe (H2020, ERC) et en début de projet : en quoi le projet pourra compter sur nous, prise en compte des bonnes pratiques, participation aux consortiums
- Début de production de données brutes : stockage sécurisé
- Elaboration des données : outils collaboratifs et de traitement
- Finalisation des corpus : mise en ligne, ouverture, référencement
- Démarche d'archivage

Ce qu'on ne fait pas

- Conception, spécifications, développement des objets numériques :
  - structuration et organisation des données
  - modèle de données
  - bases de données
  - sites Web et applications associées
- Administration des briques mises en oeuvre
  - CMS
  - applications, codes et scripts
- Evaluation scientifique : démarche agnostique
- Substitution aux porteurs scientifiques au fil du temps

Infrastructure

- Hébergement au CC-IN2P3 (Villeurbanne) : salle, réseau, sauvegarde
- 3 baies
- Infrastructure en propre
- 30 serveurs (Dell 1U Rxxx - Linux)
- 280 machines virtuelles (Linux KVM)
- 2 NAS NetApp (160 To nets utiles)
- 1 stockage distribué sécurisé (800 To nets utiles)
- 1 firewall PaloAlto 5220 (niveau 7 avec détection signatures d'attaque)
- Administration en interne (3 ETP)

Traiter - 1/3

- Mise à disposition de toutes les applications libres demandées
  - Langages de programmation et scripts : PHP, Python, Java
  - Moteurs de bases de données SQL : MySQL, PostgreSQL, PostGIS
  - Logiciels XML et autres bases de données : BaseX, eXist
  - 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 - 2/3

- Mise à disposition de certains logiciels commerciaux
- En mode jeton (utilisation sur le poste client)
  - oXygen : éditeur XML
  - ESRI ArcGIS for desktop : SIG
  - SAFE FME Desktop : outil ETL
  - Autodesk Infrastructure Design : conception 3D
  - Photoscan : photogrammétrie
  - MaxQDA : analyse de données

Traiter - 3/3

- En mode serveur (exécution sur notre infrastructure)
  - 2 serveurs FileMaker avec licences serveur
  - Kakadu : conversion JPEG2000
  - Sorenson Squeeze Server : conversion de videos
  - Business Geografic Geo : serveur WebMapping
  - ESRI ArcGIS : cartographie, analyse et diffusion de cartes
  - Photoscan : photogrammétrie
  - R Studio Server : statistiques et traitement de données
  - Sphinx : gestion d'enquêtes
- Méthodes d'accès dépendant de chaque application : interface Web, client lourd, accès SSH

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)
- Ce que cela ne fait pas : interface Web de présentation des données

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
- 160 projets - 2.7 To de données - 200 000 fichiers

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 (180), Drupal (50), Wordpress (80)
- Mais aussi : Java/Tomcat (40), BaseX (20), SolR/ElasticSearch (35)
- Des développements maison
- environ 600 sites

Diffuser - 2/3

- Quand besoin de plus d'outils et plus d'autonomie : machines virtuelles
- Fourniture de VM à façon en mode IAAS - PAAS
- 128 VM à ce jour (Debian, CentOS 7, Ubuntu, un peu de Windows)

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

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
- Participer activement à l'évolution de la plateforme du CINES (e.g. ajout de nouveaux formats comme la 3D, prise en compte du versionning)

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 - 4000 entrepôts - 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

Stocker - 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
- +5 To de données - +30 millions de fichiers

Stocker - Sharedocs 1/2

- Logiciel commercial FileRun
- Gestionnaire de fichiers entièrement Web
- Connexion possible en WebDAV
- Fonctions : partages, prévisualisation, URL de diffusion et courtes, étiquettes
- Watch-folders : couplage avec des logiciels de traitement en batch
- Nos usages : OCR, conversion de videos et d'audios, PDF2text

Stocker - Sharedocs 2/2

- Simple d'emploi - appropriation facile
- "Bureau de travail" en ligne
- Utile pour élaboration du jeu de données avant publication
- 3000 comptes
- 40 To de données - 9 millions de fichiers

Stocker - Stockage distribué et sécurisé 1/3

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
- données tièdes ou froides

Stocker - Stockage distribué et sécurisé 2/3

Fonctionnalités de la solution

- 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)
- Evolution dans le temps du niveau de la sécurité sur les données
- Accès depuis n'importe quel point de présence
- Perspective d'un réseau de données national
- Extensibilité illimitée
- Vérification de l'intégrité (signatures, bandes)
- Journalisation des accès

Stocker - Stockage distribué et sécurisé 3/3

Caractéristiques de la solution

- Matériels : serveurs et baies SAS Dell
- Logiciels : Active-Circle (commercial), annuaire LDAP, réseau VPN
- 9 serveurs sur 6 MSH et nos 2 lieux : Paris et Lyon
- Volumétrie nette utile : 800 To
- Gestion déconcentrée des comptes et des groupes : FusionDirectory (frontal Web sur OpenLDAP)
- Réseau VPN : OpenVPN auj. - L3VPN RENATER demain
- 500 To de données à ce jour

Conclusion

Qu'est-ce qu'Huma-Num ?

Pour la communauté académique de recherche française en SHS, et autour de la question des données

- une représentation de la France au niveau européen et international
- une mise en relations et une animation de la communauté
- une offre de services au service de la communauté
- une vision globale au cours de la valorisation, de l'interopérabilité et de la pérennisation des données de la recherche

Références

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