USR 3550

Quelques mots sur humanités numériques et calcul intensif dans le nuage

Digital humanities & Cloud computing ?

Thierry Chanier, MSH-LRL, Université Blaise Pascal

Réunion Cloud Computing, CRRI, Cézeaux, 26 février 2013

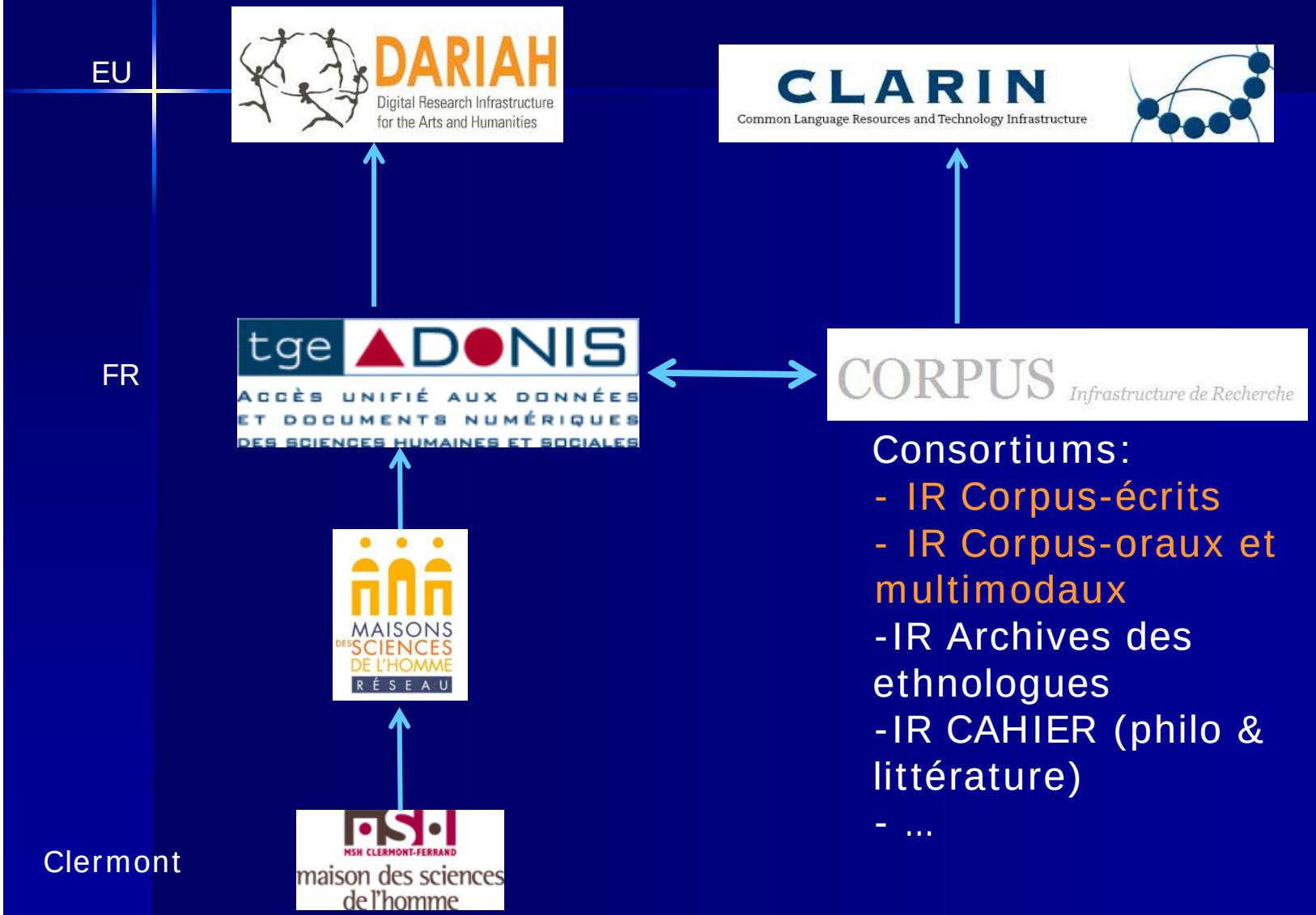
Objectifs et contenu

- Se comprendre malgré disparités disciplines
- Développement humanités numériques en Europe, France, Clermont
- Contraintes, besoins en SHS : différencier besoins et acteurs impliqués
- Enjeux à court /moyen termes vers « big data » et « cloud computing »

Union Européenne

- Just like astronomers require a virtual observatory to study the stars and other distant objects in the galaxy,
- researchers in the arts and humanities need a digital infrastructure to get access to and join together the information and the knowledge that is embedded in digital content.
- The Digital Research Infrastructure for the Arts and Humanities – DARIAH – will be such an infrastructure with a European dimension.

Organisation en France pour SHS



Consortiums:

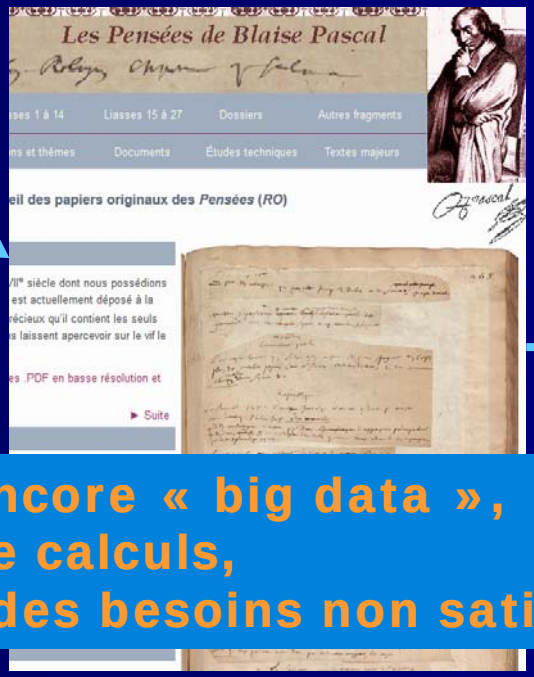
- IR Corpus-écrits
- IR Corpus-oraux et multimodaux
- IR Archives des ethnologues
- IR CAHIER (philo & littérature)
- ...

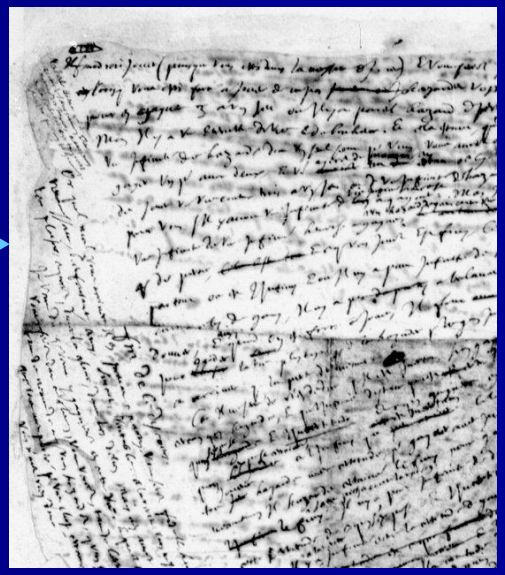
Humanités numériques et MSH Clermont: exemples avec textes et images

Serveurs MSH

En cours de développement

Numérisation fine et stockage à la BNF





Pas encore « big data »,
pas de calculs,
mais des besoins non satisfaits

Humanités numériques et MSH Clermont: exemples corpus multimodaux, calculs

Plateforme spécifique


INTELESPACE
Plateforme géomatique

LESPACE : une plateforme de recherche en géomatique à la MSH

matique spatiale est un axe de recherche commun à plusieurs laboratoires hébergés à la Ma
ciences de l'Homme de Clermont-Ferrand. Une volonté forte de développer les compétences
atiques et de mutualiser les équipements spécifiques a émergé sous l'impulsion de la MSH
toires CERAMAC, CHEC et GEOLAB. Grâce au soutien financier de la région Auvergne, du FE
niversité Blaise Pascal et du CNRS, la plateforme INTELESPACE (pour « intelligence de la do
le ») est née en 2009 concrétisant cette synergie spatiale à la MSH de Clermont-Ferrand. La
rme est animée et pilotée par Franck VAUTIER (IR CNRS) et Erwan ROUSSEL (IGE UBP).

sion et domaines de compétences

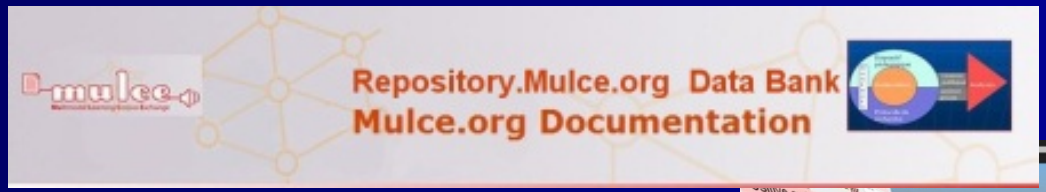
ESPACE a une mission de recherche et de développement méthodologique en géomatique.
e un soutien technique et thématique sur les trois principales étapes de la recherche à domi
le : l'acquisition de données géolocalisées, le traitement et la transformation des données et
se spatiale et l'interprétation des données.

Acquisition de données sur le terrain | Traitements et préparation des données | Analyse spatiale, interprétation, modélisation

teurs d'activités

Quelques téraoctets
Calculs sur modèles en dizaines d'heures

Serveur MSH



1 téraoctet,
Données XML,
début calculs

Textes, images, audio, vidéo

Se rapprochent « big data » ?, et des besoins non satisfaits

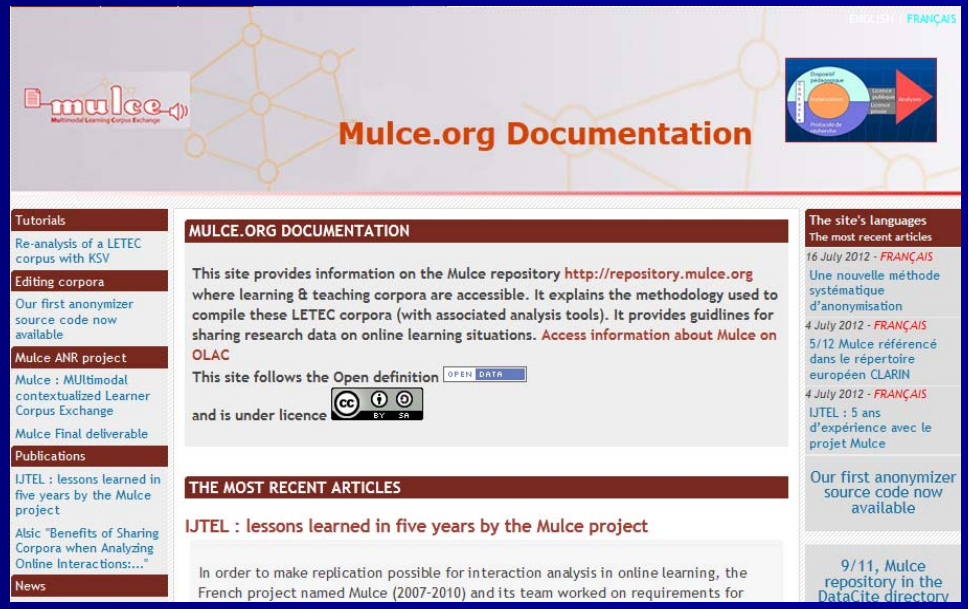
Données recherche, contraintes, et besoins en SHS (hors calcul intensif)

- Beaucoup (mais pas toutes) données collectées, structurées à la main: besoin RH qualifiées
- Formation des chercheurs sur standards
- Questions prégnantes sur éthique et droits
- Diffusion stable des données, archivage à long terme (changement formats)
- Référencement données : visibilité, moissonnage

Mulce online

Repository.Mulce.org Databank

Mulce.org Documentation

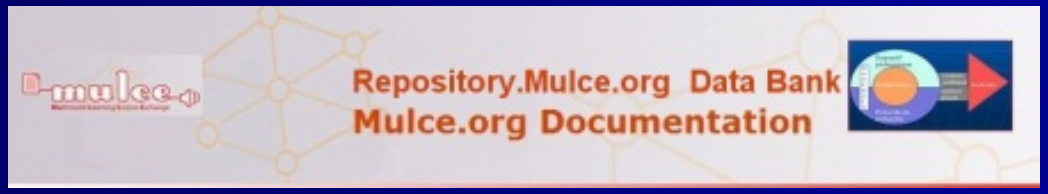


OPEN LANGUAGE ARCHIVES

CLARIN
Common Language Resources and Technology Infrastructure

Metadata in OLAC and in CLARIN

Open access, ethics and licence

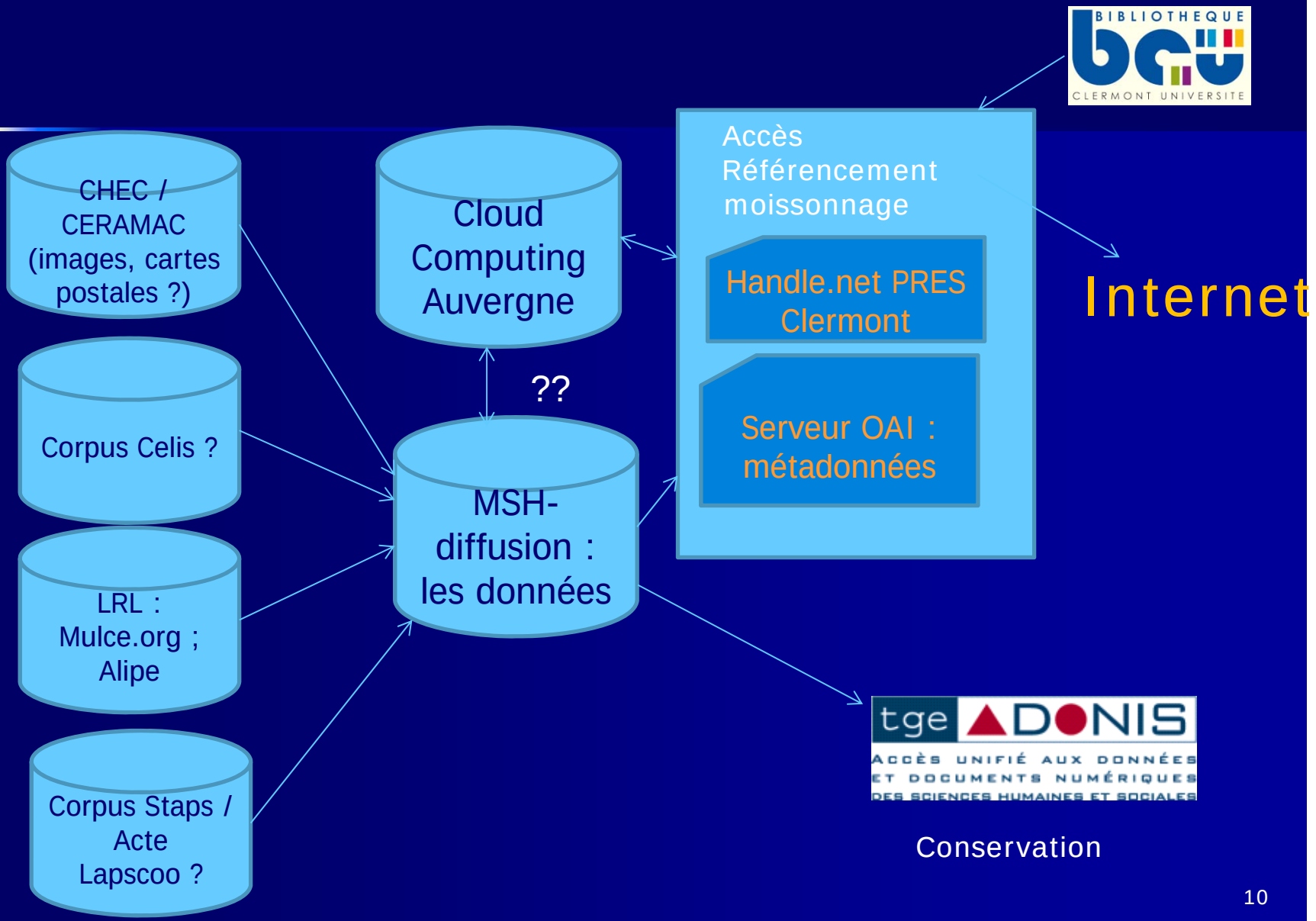


Open Data:
http://opendefinition.org/guide/

For usage: licence

For participants: Informed consent form
+
Anonymization process

MSH-PRES

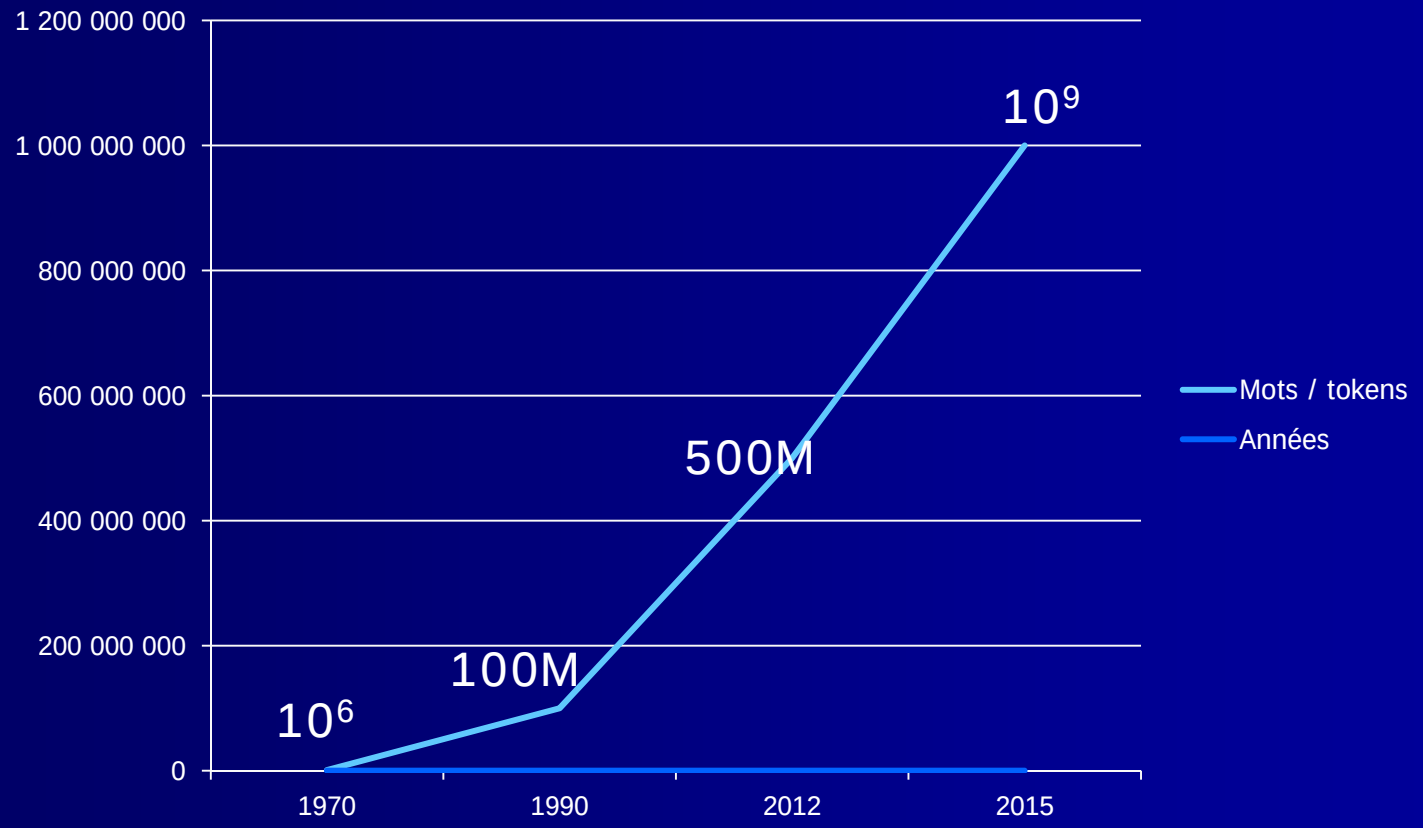

CHEC / CERAMAC (images, cartes postales ?)
Corpus Celis ?
LRL : Mulce.org ; Alipe
Corpus Staps / Acte Lapscoo ?
Cloud Computing Auvergne
??
MSH-diffusion : les données
Accès
Référencement
moissonnage
Handle.net PRES Clermont
Serveur OAI : métadonnées
Internet
BIBLIOTHEQUE
CLERMONT UNIVERSITE
tge ADONIS
ACCÈS UNIFIÉ AUX DONNÉES ET DOCUMENTS NUMÉRIQUES DES SCIENCES HUMAINES ET SOCIALES
Conservation

Exemple corpus de référence sur langue

PRÉVOIR DÉVELOPPEMENT PROCHE CLOUD COMPUTING SHS

Repères internationaux en corpus de langue

Nombre de mots / tokens



EN : Brown Corpus

EN : British BNC

NL et DE : National Corpora

Corpus sur hollandais

SoNaR – Facts and figures

- Spoken Dutch Corpus
  9-million-word corpus, incl. audio, collected between 1998 and 2003

- Need for reference corpus of contemporary written Dutch

- With funding from the Dutch-Flemish STEVIN program
  - D-Coi (pilot) project – 2005-2006
    corpus design, protocols, procedures, etc.
    54-million-word corpus
  - SoNaR project – 2008-2012
    500-million-word corpus

Corpus sur allemand

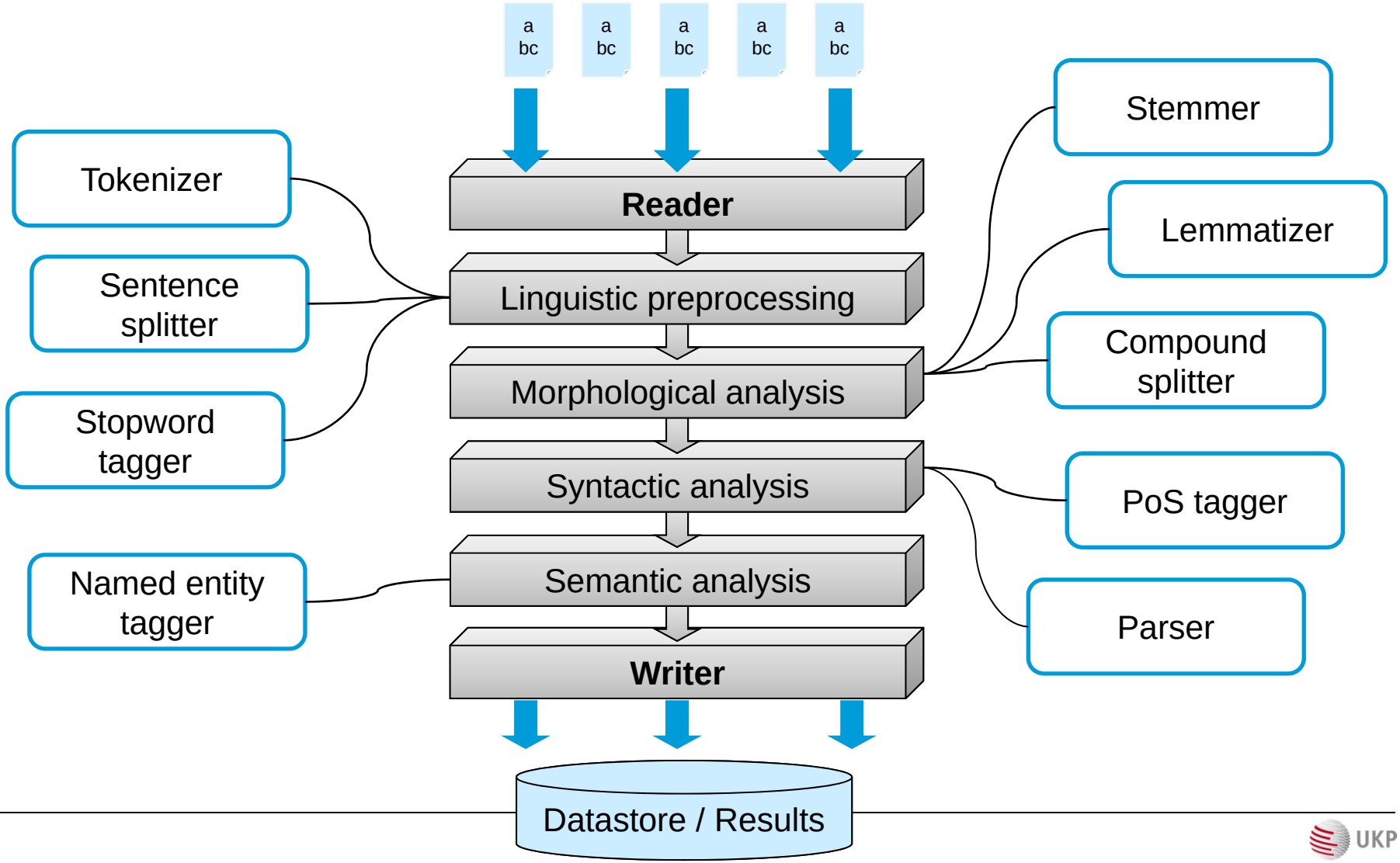
Korpora

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

- **DWDS – Kernkorpus des 20./21. Jh.**
  - 100 Millionen laufende Wörter (etwa 80.000 Dokumente)
  - ausgewogene Verteilung über Zeit und Textsorten (Schöne Literatur, Gebrauchsliteratur, Journalistische Prosa, Wissenschaftstexte)
- **DWDS – Erweitertes Korpus**
  - 1,5 Milliarden laufende Wörter (etwa 5 Millionen Dokumente): Zeitungstexte seit den 90er Jahren
  - Texte namhafter Zeitungen: Die Zeit, Süddeutsche Zeitung etc.
  - Belletristik des 21. Jh.

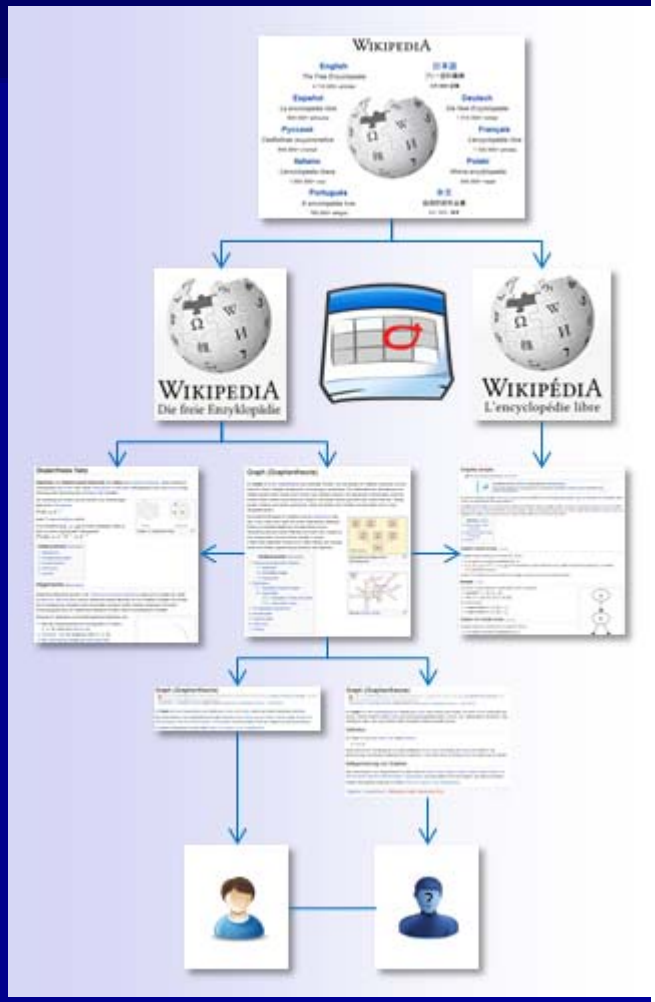
TECHNISCHE UNIVERSITÄT DARMSTADT

Pipeline Architecture



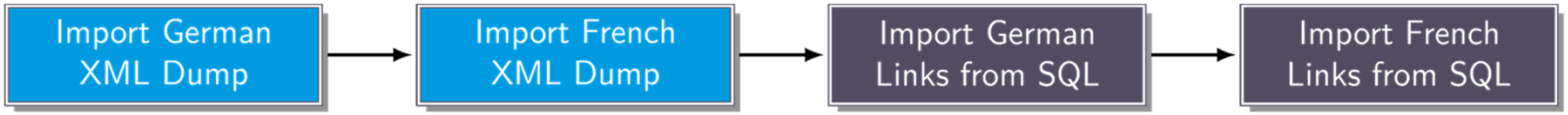
Exemple proche « big data »

- Extraction automatique et traitement Wikipedia
- 1 seul article (peintre Monet) en deux langues : FR et DE



(Goethe-Universität Frankfurt
Rüdiger Gleim, Alexander Mehler)

Import XML Dumps into the Database



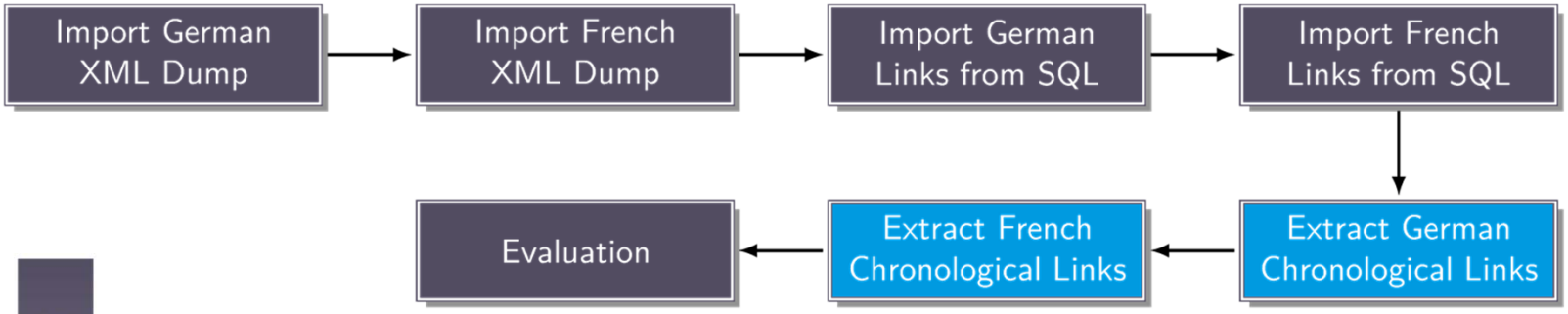
Attribute	Value (German)	Value (French)
Date of Dump	2012-12-15	2012-12-13
Pages	4 283 757	5 505 082
Revisions	101 959 908	82 296 570
XML Dump Size (7z compressed)	14,09 GB	10,01 GB
~~XML Dump Size~~	2 050,28 GB	1 319,77 GB
Text Compression Rate	2,928 %	4,187 %
Diff Compressed Revisions	92,881 %	87,367 %
Threads for Import (Text Compression)	24	24
~~Import Time~~	30,5 h	21,0 h
Average Time/Page	25,635 ms	13,76 ms

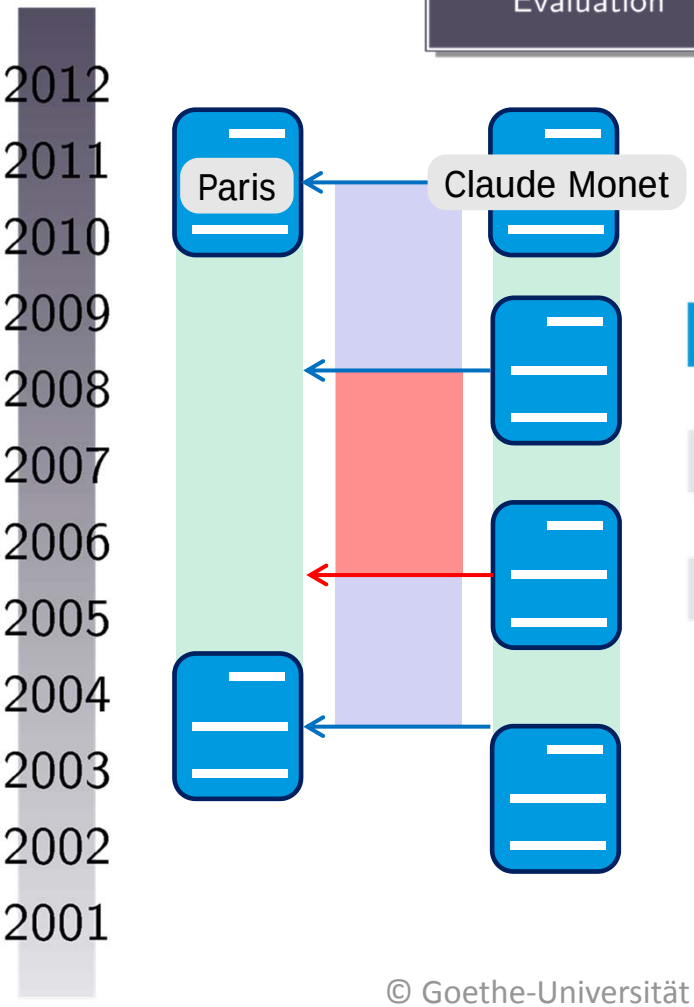
2 To

Temps calcul

```
1 // Get Instance of WikipediaRoot. Initialize Database if necessary
2 WikipediaRoot root = new WikipediaRoot_impl();
3 // Specify XML Files to import
4 File deXMLFile = new File("dewiki-20121215-pages-meta-history.xml.7z");
5 File frXMLFile = new File("frwiki-20121213-pages-meta-history.xml.7z");
6 // Import german (de) Wikipedia from specified file using up to 24 threads
7 Wikipedia deWiki = root.importWikipedia("de","2012-12-15",deXMLFile,24);
8 // Import french (fr) Wikipedia from specified file using up to 24 threads
9 Wikipedia frWiki = root.importWikipedia("fr","2012-12-13",deXMLFile,24);
```

Extract Chronological Links





Calculs sur liens
(pas encore étiquetage linguistique)

Attribute	Value (German)	Value (French)
Elapsed Time	22,98 h	38,90 h
Average Time/Page	19,31 ms	25,44 ms
Threads for Parsing	30	30
Imported Links	111 815 686	210 468 542
Extracted Links	94 161 457	99 404 533

Retour sur France et Clermont

- 2013-.., France : projet constitution corpus de référence du français contemporain (consortiums IR-corpus)
- 2013-2014, Clermont : groupe travail sous-corpus avec labos distants, dépôts données et calculs dans nuage

CORPUS *Infrastructure de Recherche*


LE WIKI DE LA LISTE CORPUS-ECRITS-NOUVCOM

Projet corpus Nouv-com

Rassemblement de corpus existants. Plusieurs membres / équipes de notre groupe de travail disposent déjà de corpus rasse
clavardage, forums, twit, etc.). Il s'agit de sélectionner tout ou partie de ces corpus déjà organisés en XML et contenant de
rassembler en une banque de corpus accessible à tous. Chaque corpus sera documenté suivant des standards (OLAC, CLARIN)
serveur mis à disposition sur un serveur national (cf. ci-dessous), faisant tourner un protocole OAI-PMH. Des membres de ce pr
d'envisager d'adjoindre cette banque au corpus global. Les membres du projet étudieront la façon dont des analyses / annotati
faire l'objet de projet de recherche financés par ailleurs - ANR, etc., la banque servant alors de point d'appui dans le montage d
TEI-nouv-com afin de voir comment à l'horizon 2013-14, les corpus de la banque pourraient être avoir une version TEI. Vous pouv
à disposition du projet vos corpus ou, si vous n'en avez pas, désirez participer à l'avancement du projet

- Coordinateur : à déterminer
- Personnes amenant des corpus : Gudrun Ledegen, Thierry Chanier, Benoît Sagot, Virginie Zampa, Achille Falaise, Georges Ant
- Autres participants : Tita Kyriacopoulou, Rachel Panckhurst
- Total : 9 personnes ; UR différentes : 8 (3 IDF, 3 Grenoble, 1 Rennes, 1 Clermont-Fd, 1 Montpellier)
- Lieu potentiel de réunion : Grenoble et Idel de France (IDF)
- Réunions et déplacements : 2 réunions en présentiel(dont 1 couplée avec celle du projet TEI nouv-com, le jour suivant) ; 2
  lieu dans Adobe Connect.

Note : Benoît et Gudrun ont émis une réserve sur le fait de participer aux travail du groupe (en dehors des réunions). Rachel n'éc