



HAL
open science

L'INGÉNIERIE DES CORPUS

Mokhtar Ben Henda

► **To cite this version:**

| Mokhtar Ben Henda. L'INGÉNIERIE DES CORPUS. Master. France. 2018. cel-01716602

HAL Id: cel-01716602

<https://cel.hal.science/cel-01716602>

Submitted on 23 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Université Bordeaux-Montaigne – MICA
Équipe E3D (Études Digitales des Données aux Dispositifs)
(EA 4426)

2017-2018

L'INGÉNIERIE DES CORPUS

Méthodes, outils et aspects normatifs

Séminaire E3D
Master Humanités Numériques

Mokhtar Ben Henda

Plan

CADRE GÉNÉRAL

ÉLÉMENTS MÉTHODOLOGIQUES

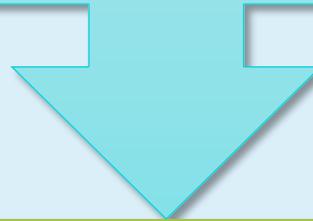
ÉLÉMENTS TECHNIQUES

Le corpus, un point d'orgue dans la recherche

- Le tournant/virage discursif des années 70s :
 - *Les objets d'études émergent discursivement ;*
 - *Analyse du discours largement applicable dans la recherche linguistique ;*
 - *Analyse des données largement applicable dans les SHS / SIC ;*
 - *Appel au langage quels que soient ses méthodes et ses domaines de recherche (Lahire 1994) ;*
 - *Avènement de la « linguistique de corpus »* (empirisme) ;*
 - *Constructivisme (radical) : démarche empirique par objet corpus ;*
 - *« Considérer que ce qui est dit et écrit médiatise une part de la réalité et y donne une prise » (Le Lay 2013)*
 - ➔ Corpus = énoncés construits qu'il faut étudier !

Convergence vers le texte

« Tous les discours étudiés en linguistique et sciences sociales prennent la forme matérielle du texte, mais la dimension discursive d'un même texte fait qu'il est le terrain de différents niveaux d'expression et de représentation, chacun pouvant être étudié [et interprété] suivant l'orientation disciplinaire du chercheur » (Comby, 2016)



La situation centrale occupée par le texte (l'intertextualité) interroge les cloisonnements disciplinaires

(Inter-multi-trans-disciplinarité).

Corpus & SHS : transversalité

Les SHS promeuvent de plus en plus la mise en place de méthodologies ancrées dans le terrain (littérature, sociolinguistique, ethnologie, didactique, sociologie, information et communication, etc.) dans un contexte marqué par :

La déconstruction des outils méthodologiques de la recherche

Une porosité des frontières disciplinaires

?

- Si le corpus apparaît aujourd'hui comme constitutif de toute recherche en SHS, sa conception et son analyse demeurent très variables selon les travaux engagés

?

- Son exploitation dans un processus de recherche prête à de nombreuses pratiques méthodologiques (sa constitution, son analyse, sa place, ses fonctions ...)

?

- Comment un chercheur en SHS peut-il s'approprier et s'approprié-t-il l'objet-corpus, depuis le « recueil » jusqu'à l'analyse, à travers l'emploi d'une méthodologie, voire la création d'une méthode ?

Définitions variées

Corpus ?

LINGUISTIQUE : « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon au langage » (Sinclair, 1996)

LINGUISTIQUE DE CORPUS « un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications » (Rastier, 2011)

SHS : « Ensemble d'éléments issus du réel, appelés "**observables**" (De Robillard), enregistrés, médiatisés par le chercheur ou préexistants (corpus littéraire, corpus oral, documents vidéos,...) qui sont recueillis puis sélectionnés et organisés pour constituer la base d'une analyse scientifique » (Le Gal, 2011)

GÉNÉRALITÉ : « ensemble de documents, artistiques ou non (textes, images, vidéos, etc.), regroupés dans une optique précise. On peut utiliser des corpus dans plusieurs domaines : études littéraires, linguistiques, scientifiques, philosophie, etc. » (Wikipedia)

Typologie

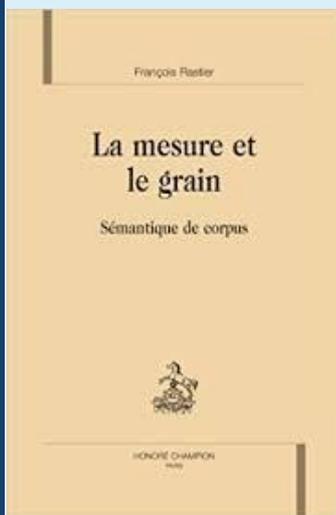
□ Deux conceptions combinées du corpus selon Rastier (2011*)

Documentaire
(logico- grammaticale)

- Ne tien pas compte de l'intégrité d'un texte
- Ne retient que des variables globales caractérisant les documents (mots, phrases), sans tenir compte de leur caractère textuel, ni de leur structure

Philologique-herméneutique
(interprétative)

- Évoque une praxéologie adaptée aux variations des tâches et des applications d'analyse
- Tient compte des rapports de texte à texte (intertextualité)



Julien Longhi, « La culture par les corpus : qualité & quantité en sémantique de corpus », Acta fabula, vol. 12, n° 8, Notes de lecture, Octobre 2011, URL : <http://www.fabula.org/acta/document6544.php>

Typologie

Corpus existant (archives)

- Masse « informe » de textes mal définis aux contours incertains auxquels on peut avoir accès.
- Cet existant dépend de conditions étrangères à l'étude, qui ne sont pas toutes connues ni maîtrisées

Corpus de référence

- Composé à partir du corpus existant, en adéquation avec l'objectif de travail.
- Clairement défini et équilibré, il fixe le point de vue de l'étude et représente le fond sur lequel on veut profiler les textes étudiés.
- En linguistique des corpus : il sert de médiation entre la **langue historique** et la **langue fonctionnelle**

Corpus d'étude

- L'ensemble des textes sur lesquels porte effectivement l'analyse.
- Délimité par les besoins de l'application.
- Subit le processus méthodologique imposé par la discipline

Corpus distingué

- Un groupe de textes du corpus d'étude que l'on veut caractériser dans leur cohésion d'ensemble, par rapport au reste du corpus d'étude.
- Un sous-corpus de travail qui varie selon les phases de l'étude et peut ne contenir que des passages pertinents du texte ou des textes étudiés

Typologie : exemple

- ❑ Frantext : base de données de textes français (1970)
- ❑ Maintenu par l'ATILF-CNRS (ex INaLF)
- ❑ Corpus d'auteurs, de périodes chronologiques, de genre)

	corpus existant	corpus de référence	corpus d'étude	corpus distingué
Etude d'Etienne Brunet (Brunet 1995)	la base Frantext de l'INaLF	350 romans entre 1830 et 1970	phrases de ces romans comportant au moins une des 165 unités lexicales retenues pour définir la thématique du sentiment	les éléments retenus dans les romans d'un romancier
Construction des profils pour l'application DECID de diffusion ciblée	textes enregistrés dans la base SPHERE de la DER d'EDF, autres textes électroniques collectés de façon centralisée.	l'ensemble des textes d'Action, en version définitive, à partir de l'année 1990 jusqu'à l'année en cours.	les textes d'Action pour une année (le cas échéant, les textes en version provisoire pour l'année suivante).	les textes d'Action du corpus d'étude, dont le rédacteur (plus exactement le responsable) est rattaché à un Département donné.

Caractéristiques

- « Le corpus n'existe pas en soi, mais dépend du positionnement théorique à partir duquel on l'envisage » (Charaudeau, 2013)
 - *Il dépend aussi :*
 - du contexte et du matériel d'étude concerné (terrain : observation sur terrain ou entretien face à face / oral, textuel ou multimodal) ;
 - du domaine concerné (inter/multi/transdisciplinarité) ;
 - de son historicité : ouvert ou clos ;
 - de sa représentativité (envergure) : corpus existant, de référence, d'étude ou distingué
 - De son exploitation : manuelle, automatisée ou les deux,
- « Tout regroupement de textes ne mérite pas le nom de corpus » (Rastier, 2011)
- « Tout ensemble de textes n'est pas un corpus » (Bommier-Pincemin, 1999)
 - « *Collection de textes avec une volonté de **cohérence*** »
 - « *Vérifie des conditions de **signifiante**, **d'acceptabilité** et **d'exploitabilité*** »

Caractéristiques (règles)

Pertinence

Règle - « Les documents retenus doivent être adéquats comme source d'information pour correspondre à l'objectif qui suscite l'analyse. (Bardin 1977, §III.I.1, p. 128)

Cohérence

Règle - représentativité d'une entité ayant un ou plusieurs caractères communs (sans trop de singularité)

Représentativité

Règle - un échantillonnage rigoureux (équilibré/diversité maximale) dont les résultats sont généralisables à tout l'ensemble (recherche de diversité maximale)

Régularité

Règle - non sélectivité : ne pas permettre d'exceptions pour éviter des écarts d'analyse (manques, excès, éléments étrangers)

Complétude

Règle - un niveau de détail adapté aux besoins de l'analyse

Homogénéité

Règle - toutes les grandeurs recensées [variations] sont des quantités de même nature

Volume

Règle - important pour des analyses statistiques voulues significatives

ÉLÉMENTS MÉTHODOLOGIQUES

- Les corpus dans la pratique de la recherche
- Approches méthodologiques : quantitative Vs Qualitative
- Interdisciplinarité SIC/SHS

Recherche fondamentale et/ou appliquée ?

- ❑ Historiquement deux courants rivaux (Monde académique vs monde industriel) ;
- ❑ Aujourd'hui, la ligne de démarcation est encore très floue (OCDE, 2003)
- ❑ Consensus sur une nouvelle conception de la recherche fondamentale (OCDE, 2003*) ;
- ❑ La Recherche fondamentale consiste en « *des travaux expérimentaux ou théoriques entrepris essentiellement en vue d'acquérir de nouvelles connaissances sur les fondements de phénomènes ou de faits observables, sans qu'aucune application ou utilisation pratiques ne soient directement prévues.* » (Journal officiel 2006/C 323/01 du 30/12/2006)
 - *Fondée à la fois sur la curiosité pure, sans aucune application en vue, et la recherche inspirée par des applications éventuelles*
 - *Couvre l'ensemble des types de recherche nécessaire au développement d'un corpus cohérent de savoir pouvant se traduire en avancées socioéconomiques.*

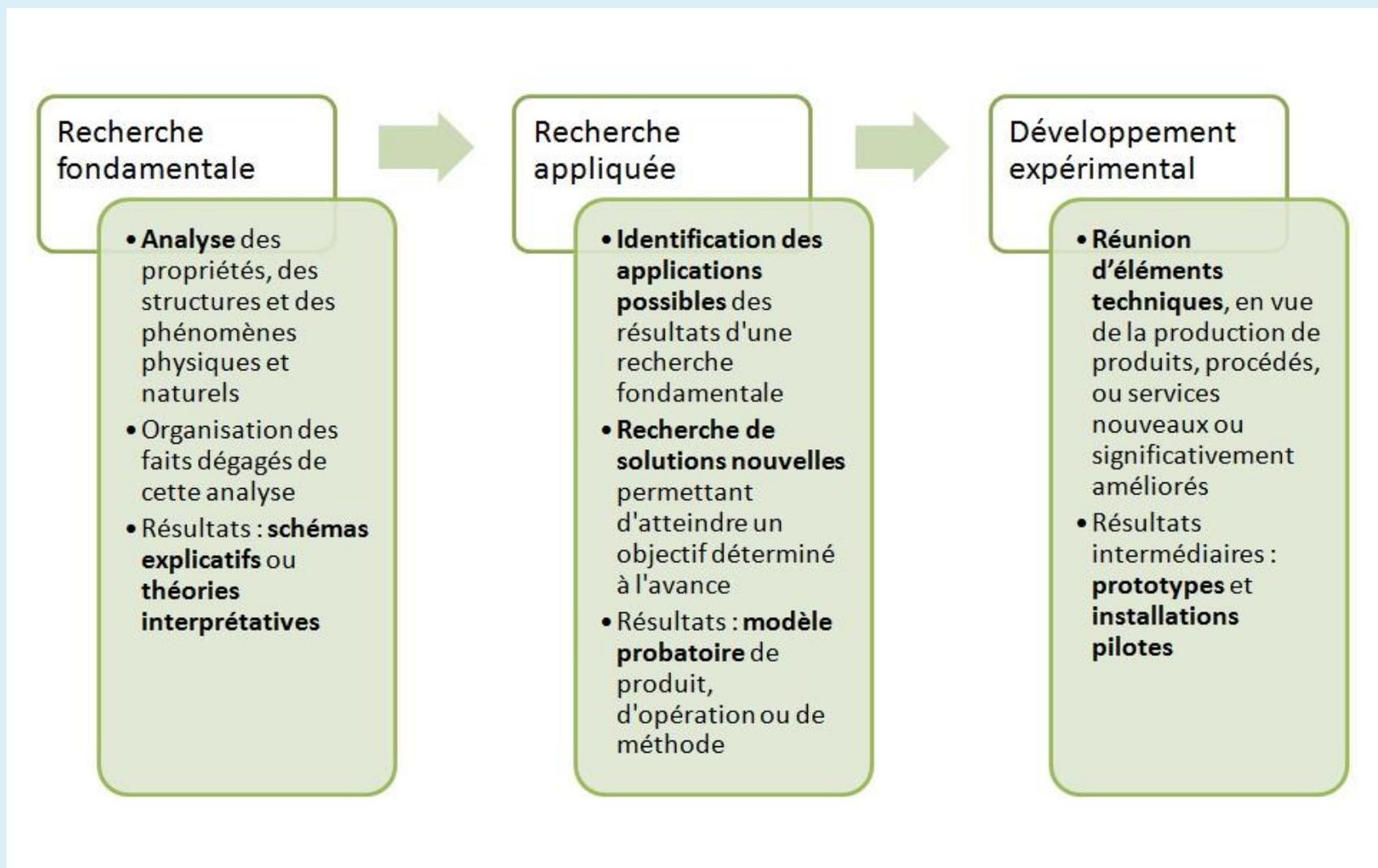


OECD. Gouvernance de la recherche publique : Vers de meilleures pratiques. OECD Publishing; 2003.



Collectif. Manuel de Frascati 2015: Lignes directrices pour le recueil et la communication des données sur la recherche et le développement expérimental. OECD; 2016

Recherche fondamentale et/ou appliquée ?



Corpus dans le processus méthodologiques de la Rech.

❑ Le choix s'opère au moment de l'élaboration des hypothèses et des questions de recherche

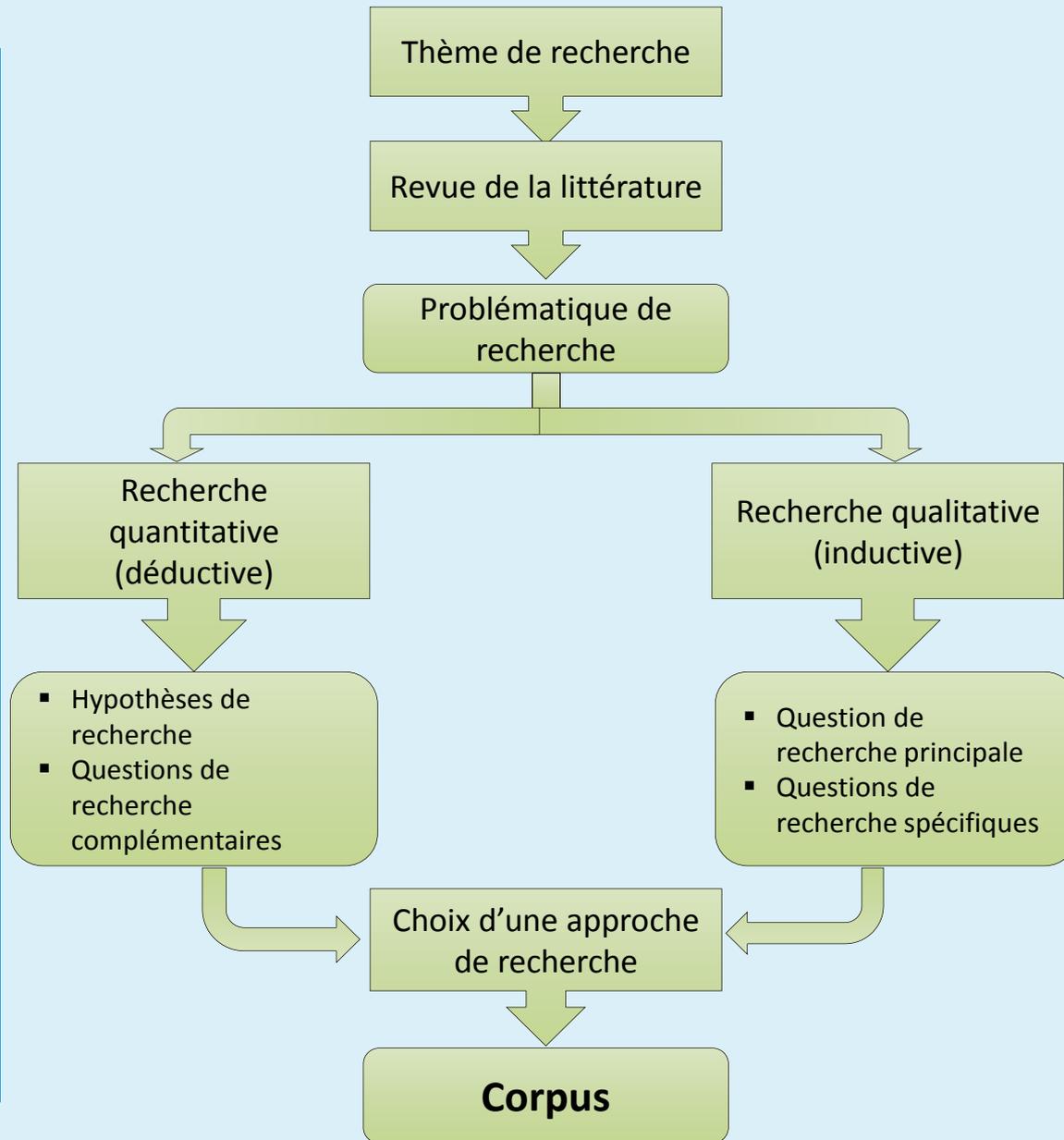
❑ Approche quantitative ou déductive

- Cohérence avec les hypothèses ou théories identifiées en prémisses par le chercheur
- Précise la formulation des hypothèses
- De l'anticipation par rapport aux résultats.

❑ Approche qualitative ou inductive (théorie de l'induction):

- Phénoménologie : une vision du monde où la réalité est multiple
- Paradigme constructiviste
- Aboutir à une idée par généralisation et non à partir d'hypothèses préétablies.

❑ Approche mixte ?



Facteurs tangibles

- La faisabilité d'une étude se construit en fonction :
 - *des objectifs fixés au départ ;*
 - *des choix méthodologiques ;*
 - *du terrain (périmètre) à étudier ;*
 - *des méthodes d'enquête/analyse ;*
 - *du positionnement du chercheur par rapport à son corpus d'étude ;*
 - *de sa capacité à adapter ses choix aux obstacles,*

- Les choix méthodologiques impactent :
 - *Le recueil des données ;*
 - *L'échantillonnage et la représentativité ;*
 - *L'analyse des données ;*
 - *L'exploitation des résultats.*

Grandes questions pour le chercheur

La nature du corpus

- Sa sélection, sa construction, son organisation, sa hiérarchisation

Le conditionnement des pratiques méthodologiques

- Les choix effectués quant à la constitution du corpus et réciproquement, cadre théorique et/ou pratique préalable et adaptabilité à la réalité du terrain

Le positionnement épistémologique du chercheur

- L'implication du sujet-chercheur dans son corpus : distanciation par rapport à l'objet, explicitation des choix méthodologiques

La finalité sociale et scientifique de la recherche

- L'articulation de cette méthode avec le rôle social de la recherche

Risque majeur pour le chercheur

« Les dangers liés à une mauvaise gestion des finalités d'une recherche consisteraient par exemple à ne développer à partir de son corpus que les caractéristiques, les tendances corroborant la/les hypothèses que l'on s'est forgées, à n'y « trouver » que ce que l'on souhaite par avance, laissant de côté les résultats des analyses des cas non conformes aux hypothèses, ceux que le chercheur ne comprend pas et/ou qui contredisent ses thèses »

(Le gal, 2011)

Approche quantitative (déductive)

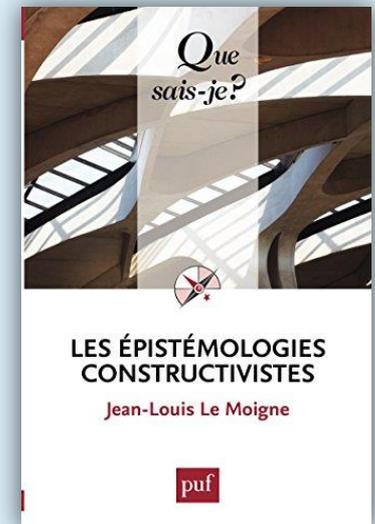
- S'effectue sur des données recueillies (les variables) mesurables :
 - *Données d'enquête (sondages, questionnaires, données statistiques préexistantes) ;*
 - *Données produites par encodage numérique (documents d'archives, dossiers administratifs, sources sonores ou visuelles) ;*

- A pour finalité :
 - *D'étayer une théorie/hypothèse ;*
 - *D'accompagner un raisonnement de recherche dans sa démarche empirique ;*
 - *Déterminer la relation générale entre un énoncé (variable) et une autre variable indépendante dans un corpus donné ;*
 - *Obtenir des informations qui peuvent être déduites d'un échantillon puis généralisées à de larges populations d'unités (corpus existant).*

Approche qualitative (inductive)

- Utilise des techniques variées :
 - *entretiens individuels semi-directifs approfondis ;*
 - *discussions de groupe (« focus groups ») ;*
 - *analyse de contenu ;*
 - *observation participative ;*
 - *histoires ;*
 - *récits de vie...*

- Se ressourcent dans le courant constructiviste :
 - *Le constructivisme brise la conception ontologique (métaphysique) de la réalité qui existerait en elle-même, indépendamment de nous ;*
 - *Le constructivisme **radical** : la réalité n'existe pas en dehors de notre imagination. Elle raisonne plutôt en termes d'interactions complexes entre les usagers entre eux et avec l'environnement socialement et historiquement construit (y compris les dispositifs sociotechniques).*



Moigne J-LL. Les épistémologies constructivistes: « Que sais-je ? » n° 2969. Presses Universitaires de France; 2012.

Approche qualitative (inductive)

❑ Les dix conditions d'une analyse qualitative (Paillé, 2011*)

Approche terrain

éviter les hypothèses théoriques → alternance collecte/analyse et ancrage dans les données empiriques du contexte

Logique de proximité

avec les phénomènes observés, les acteurs, le contexte et le chercheur

Travail de l'esprit

seul l'esprit humain peut extraire le plus de sens d'une donnée brute

Quête de sens

un mot n'a pas de valeur absolue, son sens est de l'ordre d'une transaction issue d'une interprétation

Pratique artisanale

« artiste », « artisan » et « technocrate »

Orientation clinique

centrée sur le cas comme phénomène singulier

Visée pragmatiste

obéir plus aux règles d'une pratique qu'aux règles d'une science

Optique interprétative

fixation sur les objets au détriment des méthodes

Finalité narrative

nature dialogique : « une chaîne de paroles, par laquelle se constitue une communauté de culture et par laquelle cette communauté s'interprète elle-même par voie narrative » (Ricoeur, 1986)

Démarche explicite

Mais pas trop formaliste, avec des opérations repérables et des règles de décision manifestes et constamment soumises à révision



Un dilemme méthodologique !

- ❑ Approche empirico-inductive ou hypothético-déductive ?
- ❑ **Inductive** ? priorité aux pratiques réelles, au contexte social :
 - *Ce n'est qu'une fois les faits observés, les données recueillies, que les concepts théoriques sont introduits pour expliquer et interpréter les phénomènes examinés ;*
 - *« les chercheurs tentent de développer une compréhension des phénomènes à partir d'un tissu de données, plutôt que de recueillir des données pour évaluer un modèle théorique préconçu ou des hypothèses à priori » (Blanchet, 2000 : 30)**
 - *La réalité n'existe pas en soi, elle est socialement construite par un ensemble d'opérations (praxéologie), des processus sociaux et d'informations sociales dont nous n'avons pas nécessairement conscience ;*
 - *La qualité de la recherche, selon une conception constructiviste, dépend de la capacité du chercheur à adapter son analyse au vu des résultats, et de prendre conscience de cette dépendance entre méthode et résultats. Ce dernier principe correspond à celui de la « récursivité de la connaissance » (Mucchielli, 2006)*

Un dilemme méthodologique !

- *Les objets techniques sont ce que Bruno Latour (1991) appelle « hybride socio-technique »,*
- *C'est ce que Michel Foucault appelle « dispositifs » : « un ensemble résolument hétérogène, comportant des discours, des institutions, des aménagements architecturaux, des décisions réglementaires, des lois, des mesures administratives, des énoncés scientifiques, des propositions philanthropiques, bref : du dit aussi bien que du non-dit » (Foucault, 1977 : 299)*
- *« La connaissance (...) ne peut être le résultat de la réception passive, mais constitue au contraire le produit de l'activité d'un sujet » (Von Glasersfeld, 1988)*
- *« La conception constructiviste du monde est potentiellement libératrice, au sens où elle permet à ceux qui l'adoptent d'exploiter leur potentiel créatif » (Segal, 1990).*

Un dilemme méthodologique !

❑ **Déductive ?** accorder sa priorité aux expériences :

- *L'observateur ne peut qu'introduire des « biais », ou au minimum une distorsion, dans la réalité observée :*
 - ➔ danger pour une observation « pure » (paradoxe de l'observateur neutre)
- *Les connaissances découlent directement et exclusivement de l'observation de l'expérience ;*
- *Très pratiqué dans les sciences de la nature, en médecine, physique ou chimie :*
 - La médecine est l'adepte principal de la pratique fondée sur les évidences, son influence se traduit par un soutien prépondérant de la méthode quantitative de recherche ;
 - « considérée habituellement comme synonyme du positivisme, engagée envers la découverte de lois universelles, utilisant une théorie neutre d'observation basée sur la mesure » (Hunt, 2011*)

Regard sur la pratique qualitative en SHS

« l'analyse qualitative comme théorie et comme pratique est pratiquement invisible au sein de la sociologie française et même, ce qui est pire, au sein des comptes rendus d'enquête et, pire encore, au sein des ouvrages méthodologiques sur la conduite de l'enquête » (Paillé, 2011)

- ❑ Le paradoxe : on ne peut s'imprégner d'un corpus de recherche au point que l'interprétation se donne d'elle-même (solidité, justesse et validité) ;
- ❑ Ce serait adopter une optique positiviste (objectivation/certitude) que de penser que les opérations d'analyse sont productrices en soi du sens ?
- ❑ Des recommandations ?
 - *Dissiper le flou méthodologique autour des opérations d'analyse et d'interprétation des matériaux étudiés ;*
 - *Prolonger les opérations d'analyse vers la mise à jour du sens ;*
 - *Optimiser l'usage des logiciels dans les méthodes d'analyse ;*
 - *S'aligner sur des normes de travail validées (acceptations sociale et technique)*

La solution mixte : la voie du compromis !

- ❑ Toute vision hiérarchique ou exclusive de ces deux approches est litigieuse (Lafflame, 2007*)

- ❑ « Une procédure mixte est justifiable ... dans un travail interprétatif raisonné et hautement réflexif » (Franceschini, dans Mahmoudian ; Mondanda, 1998).

- ❑ Les deux approches sont essentielles aux SHS :
 - *La quantification peut conduire à la vérité, l'observation rigoureuse en est aussi capable ;*
 - *Elles ont les mêmes impératifs de la rigueur scientifique (triangulation) ;*
 - *Elles questionnent la représentativité de leurs résultats ;*
 - *Elles peuvent être complémentaires en :*
 - Vérifiant différemment une hypothèse comparable ;
 - Ouvrant à la recherche des univers dissemblables ;
 - Vérifiant ce qui a été découvert avec l'approche opposée, ex. :

Le quantitatif en SHS (Brest)

- ❑ « Depuis les années 1990 les chercheurs en sciences humaines et sociales ont multiplié de nouvelles façons de produire et de traiter les données allant de paire avec l'émergence, la généralisation d'outils informatiques puis d'internet.
- ❑ Les **corpus de sciences humaines et sociales** ont la particularité d'une certaine **hétérogénéité** et une des difficultés réside dans le fait de les mettre en lien entre elles.
- ❑ Il s'agit de croiser nos expérimentations en **décloisonnant** les préoccupations **disciplinaires** pour prendre au sérieux d'emblée dans nos protocoles de recherche le triptyque **corpus/outils d'analyse/traitements** qui déterminent les conditions de production des données ».

<http://outiquanti.hypotheses.org/>

CYCLE DE RENCONTRES ET D'ÉCHANGES
TRAITEMENTS ET ANALYSES
DE DONNÉES QUANTITATIVES EN SHS

otheses.org/

otheses.org/

https://outiquanti.hypotheses.org/

Université de Bretagne Occidentale

Illustration : Dusan Trajkovic

Salle C219
de 14 h à 16 h

SÉMINAIRE
WEB SCRAPING,
EXTRACTION DE CONTENUS DE SITES
INTERNET ET ANALYSES
Jeudi 25 janvier 2018

Jean-Baptiste PRESSAC CRBC - CNRS
Bénédicte HAVARD DUCLOS LABERS - UBO

Contact et inscription
laurent.mell@univ-brest.fr

NAIRE
SEAU :
LAGES
r 2018

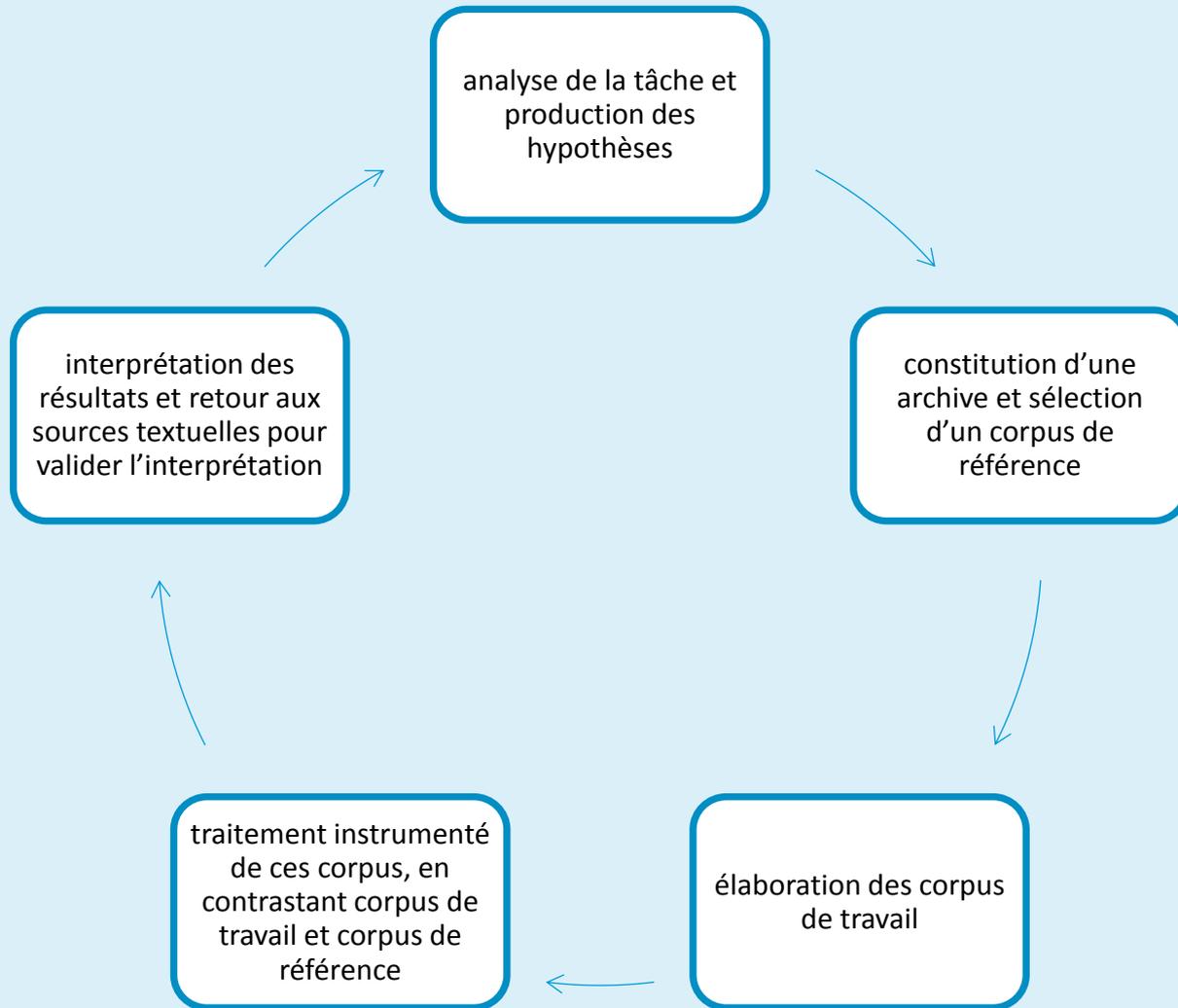
et inscription
@univ-brest.fr

IBSHS BREST
FACULTÉ DES LETTRES & SCIENCES HUMAINES

LABERS
LABORATOIRE D'ANALYSE ET DE TRAITEMENT DES DONNÉES QUANTITATIVES EN SHS

CRBC BREST

Méthode François Rastier



Consignes méthodologiques en SHS

■ Prendre du recul et revoir sa méthodologie au fur et à mesure des résultats (pertinence)

■ Retenir qu'une science n'est pas figée et qu'elle s'enrichit dans la confrontation aux autres sciences (mutualité)

■ Tenir compte de l'approche interdisciplinaire pour adopter les différents points de vue des autres champs du savoir

■ Respecte les méthodes propres à sa propre discipline de façon à valider ses résultats

■ Cultiver sa spécificités tout en gardant l'esprit ouvert.
(Mucchielli, 2006)

La trans-multi-interdisciplinarité en SHS

□ Quelques citations :

- *Claude Lévi-Strauss (1958 ; 1962) invitait, au nom de la méthode du bricolage, à établir des connexions entre l'anthropologie, la linguistique, la littérature, l'art, la psychologie, le droit, la religion, etc.*
- *Edgar Morin incite, au-delà même de la transdisciplinarité, à « écologiser les disciplines » en tenant compte de « tout ce qui est contextuel y compris des conditions culturelles et sociales » et en adoptant parfois un point de vue « métadisciplinaire » ;*
- *« Le renoncement à la complétude et à l'exhaustivité est une condition de la connaissance de la connaissance » (Morin, 1986) ;*
- *« Toute certitude fondamentale et toute croyance en un achèvement de la connaissance doivent être éliminées à jamais » (Morin, 1991) ;*
- *« Il est possible de construire une transversalité entre plusieurs disciplines, à condition de le faire d'un lieu géométrique, d'un lieu disciplinaire, faute de quoi il n'y aurait plus de validation possible du savoir » (Charaudeau, 1997, P. 12 - 13) ;*

Les SIC : le 'tourment' épistémologique

- ❑ Les SIC s'inscrivent très tôt dans une double voie :

- ❑ L'interdisciplinarité :
 - *Norbert Wiener (Cybernétique), Shanon et Weaver (Science de la transmission et du traitement du signal, Herbert Simon (Processus de décision opérationnelle) reconnus parmi les fondateurs des SIC ;*
 - *En 1993, le Conseil National des Universités (CNU) définit les Sciences de l'Information et de la Communication (SIC) comme une science interdisciplinaire ;*
 - *« une science d'adjonction, c'est-à-dire une science inter, trans et pluridisciplinaire » (Miège, 2004)*

- ❑ L'épistémologie constructiviste :
 - *« Nous construisons la réalité, nous l'inventons plus que nous la découvrons : « avec le constructivisme (...) toute prétendue réalité est (...) la construction de ceux qui croient l'avoir découverte » (Watzlawick, 1988) ;*
 - *Ces « nouvelles sciences » vont ainsi être à l'origine du renouvellement du constructivisme et du qualificatif « radical » qui lui est désormais accolé. Elles vont en effet impulser une véritable « entreprise de reconstruction épistémologique, adaptant la science contemporaine à la production de connaissances-processus plutôt qu'à la découverte de savoirs stables » (Le Moigne, 2001).*

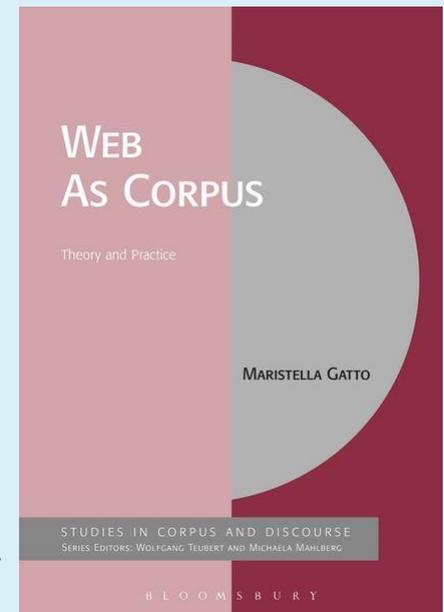


Éléments techniques

- TIC et corpus numériques
- Les corpus dans l'histoire des Humanités numériques
- Normes de balisage de corpus
- La TEI, une solution au cœur des corpus numériques
- Démo d'exemples pratiques

Corpus : éléments techniques

- ❑ Internet : notion de « Web as Corpus » [La Toile en tant que Corpus] ;
- ❑ *WebBootCaT* : un exemple d'outil de siphonage permettant la création de corpus ;
- ❑ Toute récolte numérique nécessite beaucoup de travail avant de pouvoir prétendre au nom de corpus.

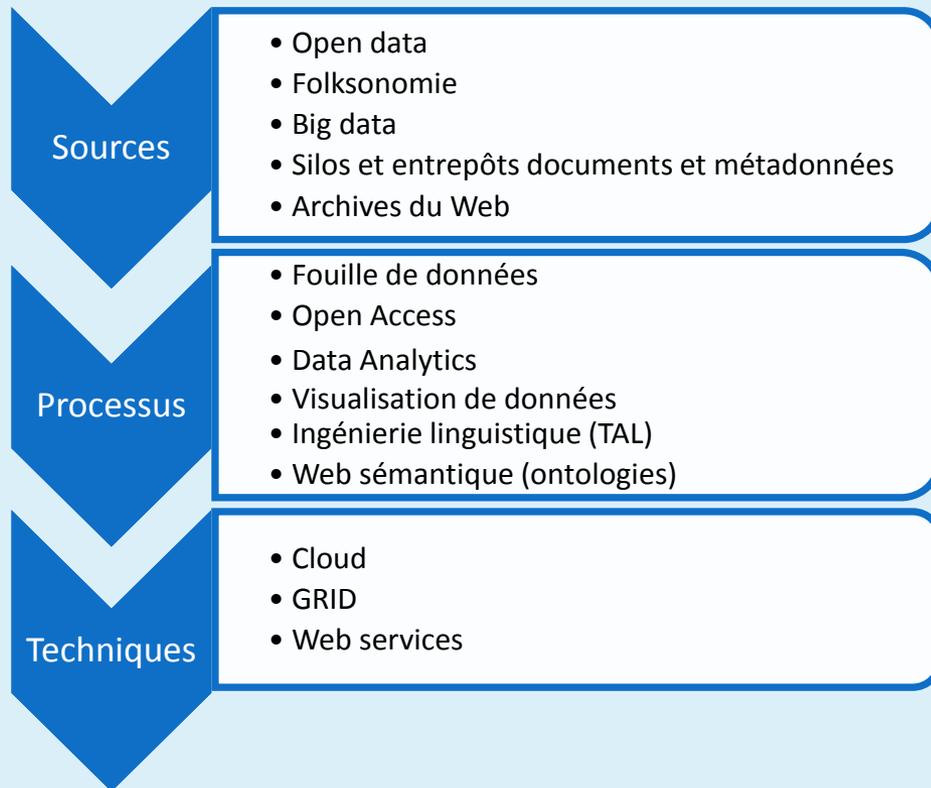


Gatto M. *Web As Corpus: Theory and Practice*.
A&C Black; 2014. 255 p.

Nouveautés : sources, Processus, techniques

❑ Le numérique des données :

- *Plus de données observables ;*
- *Plus d'alternatives de constitution de corpus ;*
- *Au cœur des Humanités numériques*



« Les humanités numériques désignent un dialogue interdisciplinaire sur la dimension numérique des recherches en sciences humaines et sociales, au niveau des outils, des méthodes, des objets d'études et des modes de communication » (Marin Dacos et Pierre Mounier, 2014)

Nouveaux enjeux, nouveaux défis

- ❑ Tenir compte des dangers de la « technologisation » liés aux corpus :
 - *Des peurs épistémologiques :*
 - Peur de non représentativité (mythe de l'universalité des phénomènes) ;
 - Peur de non exhaustivité (mythe de l'exhaustivité de l'analyse),
 - *Nouveaux supports de données :*
 - Le papier connecté (nouvelle frontière entre numérique et papier) ;
 - Réalité virtuelle / réalité augmentée (données de synthèse) ;
 - *Complexité des ressources scientifiques :*
 - Nouveaux styles artistiques (eg. NetArt) ;
 - Nouveaux genres littéraires (eg. Dark Romance, SteamPunk) ;

Nouveaux enjeux, nouveaux défis

Les genres littéraires classiques

Si vous avez appris vos cours de français à l'époque, cela devrait vous rappeler des choses...

La poésie

Haïku
Ode
Fable
Chanson
...

L'argumentatif

Pamphlet
Essai
...

Le théâtre

Tragique
Comédie
One man show
...

Le narratif

Roman
Nouvelle
Biographie
Conte
...

Le graphique

Roman graphique
Bande dessinée
Manga

L'épistolaire

Lettre
Epître

17 genres narratifs à la mode

Les genres narratifs s'appliquent au roman, mais aussi aux romans graphiques / BD.



La biographie

L'autofiction



Le roman d'amour

La Chicklit
La dark romance
La romantic fantasy



Le roman policier

Le true crime



La science fiction

Anticipation
Steampunk
Cyberpunk
Space Opera
Utopie / Dystopie
Uchronie



La fantasy

La dark fantasy
L'urbaine fantasy
L'heroic fantasy
La space fantasy



Le fantastique

Le roman gothique
Le slatterpunk

Nouveaux défis, nouvelles mesures

- ❑ Trouver des nouvelles mesures de gestion de corpus numériques conformes aux spécifications RAID :
 - *Réutilisables ;*
 - *Adaptables ;*
 - *Interopérables ;*
 - *Durables.*

- ❑ Produire des formes de normalisation et de standardisation (des référentiels normatifs communs) :
 - *qui rendent les données de la recherche intelligibles, compatibles et exploitables entre elles (linked data)*
 - *qui permettent de préserver, exploiter, produire et diffuser de données constitutives d'un patrimoine culturel (corpus, archives, bases de données, systèmes documentaires, etc.)*

Référentiel normatifs pour corpus HN

- **ADeX** – Archaeological Data eXchange: Standard for the exchange of archaeological subject data
- **CEI** – Charters Encoding Initiative: Standard for encoding historical charters
- **TEI** – Text Encoding Initiative: Standard for encoding textual data
- **EpiDoc** – Epigraphic Documents: Standard for encoding epigraphic inscriptions in TEI XML
- **CIDOC-CRM** – CIDOC Conceptual Reference Model: Ontology for cultural heritage data
- **MEI** – Music Encoding Initiative: Standard for encoding music scores

La TEI, au cœur des HN

- ❑ Les promoteurs des HD proposent de vulgariser certaines approches et règles de bonnes pratiques existantes ;
- ❑ L'une des premières « bonnes pratiques » est l'application informatique d'un schéma d'encodage normalisé à des textes numériques en SHS ;
- ❑ La TEI (Text Encoding Initiative) est ainsi née ;
- ❑ Définies en 1987, les premières directives de la TEI (TEI Guidelines) ont été publiées en Mai 1994 par Lu Burnard
 - *« Nous travaillons sur le texte, qui représente un discours, raconte une histoire, et tâchons d'expliquer ces histoires, ces contes, ces représentations » (Burnard, 2012)*
 - *« Nous sommes des experts de la maïeutique du texte, et c'est précisément ce qui définit la contribution des sciences humaines et sociales à l'élaboration du Web sémantique » (Burnard, 2012).*

La TEI, rencontre historique avec les HN

- ❑ 1960-1980 : *literacy and linguistic computing*
 - *La statistique des textes (Occurrences, concordances, fréquences des mots dans les textes (Index Thomisticum / Brown Corpus) ;*
- ❑ 1980-1994 : *text encoding (langages de balisage) ;*
 - *Représentation numérique des ressources du monde réel (livres, objets d'art...)*
 - *Émergence des langages de documents structurés (GML, ODA, SGML, TEI)*
- ❑ 1994- : *Les Humanités digitales ;*
 - *Bibliothèques numériques, GRID, Cloud, Folksnomie ;*

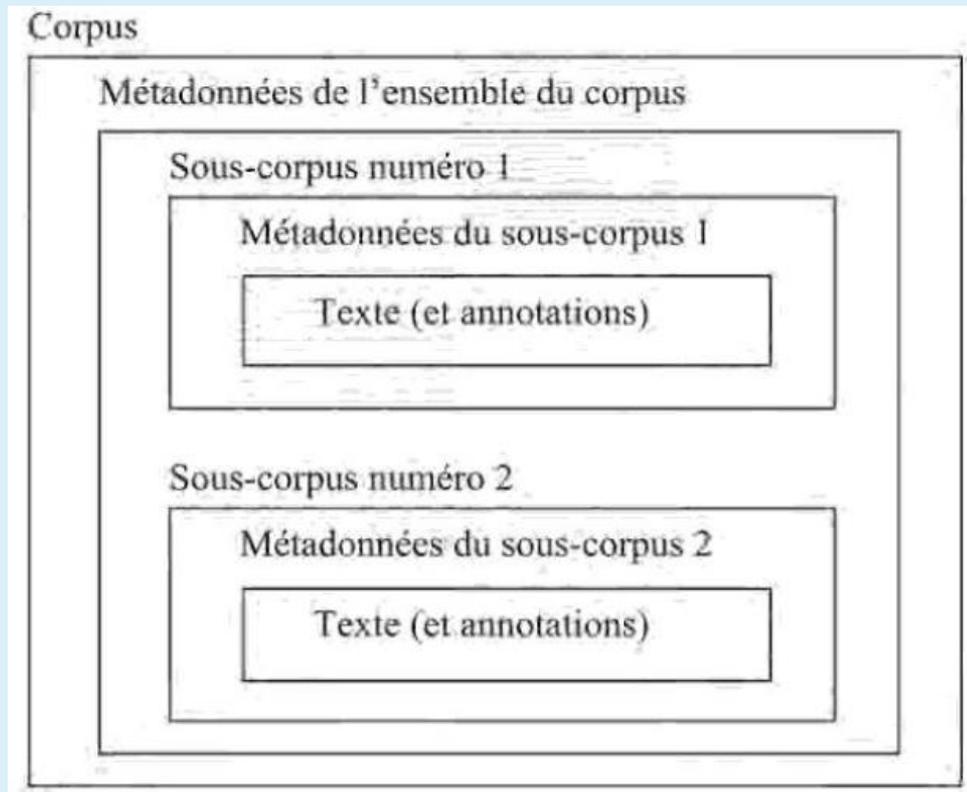
Les « humanités délivrées » Cultures parlées, visuelles et écrites, réinventées hors du livre 1-2 octobre 2013, Amphimax 414, Université de Lausanne

TEI : heuristique de construction du sens

- ❑ La TEI est une recension aussi large que possible des pratiques d'encodages et d'annotation de textes, et propose une normalisation des balises pour tous ces besoins et une formalisation de leur définition ;
- ❑ Méthode :
 - *on commence par se mettre d'accord sur la nature des faits à représenter, puis on définit les solutions d'exprimer ce consensus ;*
 - *Introduire dans le texte, au moyen d'un ensemble conventionnel d'étiquettes lisibles, des indicateurs de caractéristiques textuelles (annotation) ;*
- ❑ L'annotation :
 - *Tout corpus est intrinsèquement annoté ;*
 - *Un acte linguistique, interprétatif ;*
 - *Une structuration très simple peut supporter plusieurs niveaux d'annotation complexe*

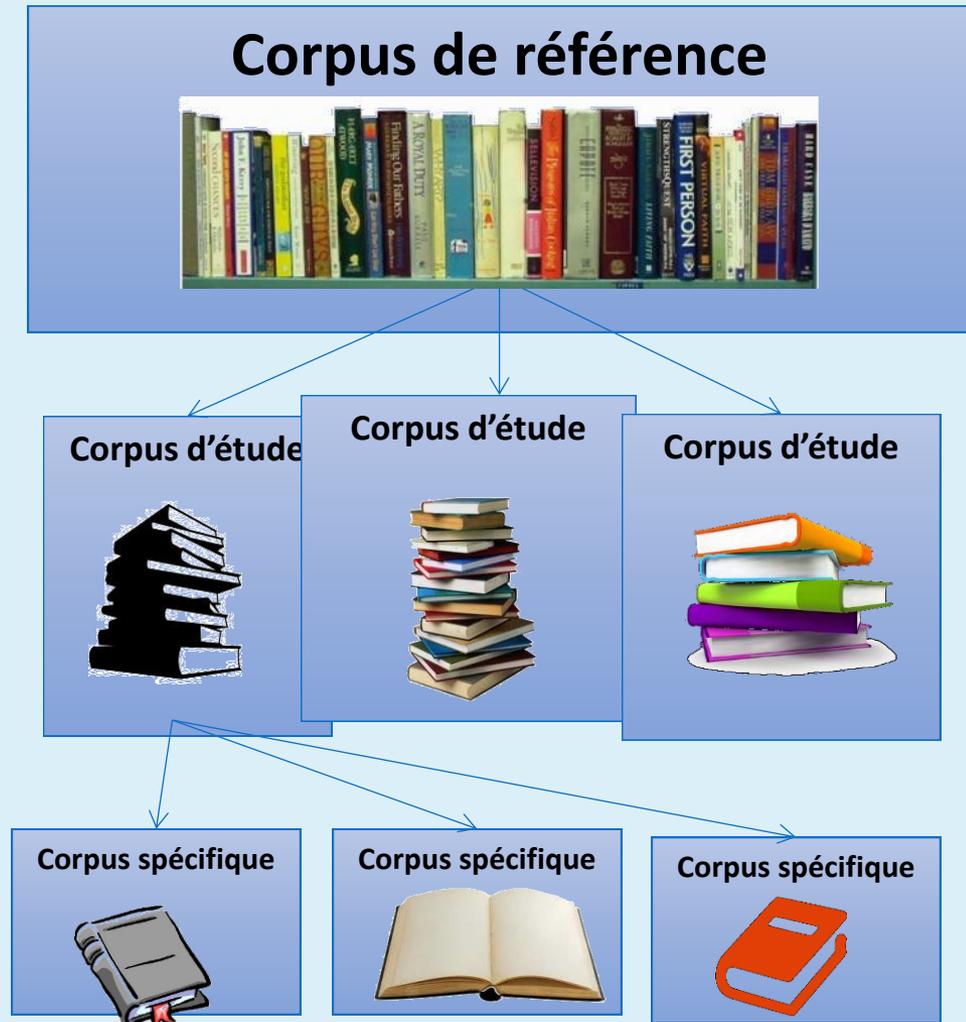
TEI : structuration générale de corpus

- ❑ Marquage des caractéristiques structurelles sous-jacentes d'un texte (phrases, paragraphes, sections, notes de bas de page, etc.).



TEI : logique hiérarchique

- ❑ Trois niveaux de représentation
 - *Niveau 1 : Tous les corpus de référence disposent d'un minimum de structure commune (Core TAG SET)*
 - *Niveau 2 : Les corpus de référence peuvent avoir des sous-corpus d'étude par genres ou types disposant d'un minimum de points communs (Base TAG SET)*
 - *Niveau 3 : Chaque sous-corpus d'étude peut avoir de corpus spécifiques*



TEI: logique structurelle

- ❑ Structure minimaliste (inspiration livresque)

“front”



“body”

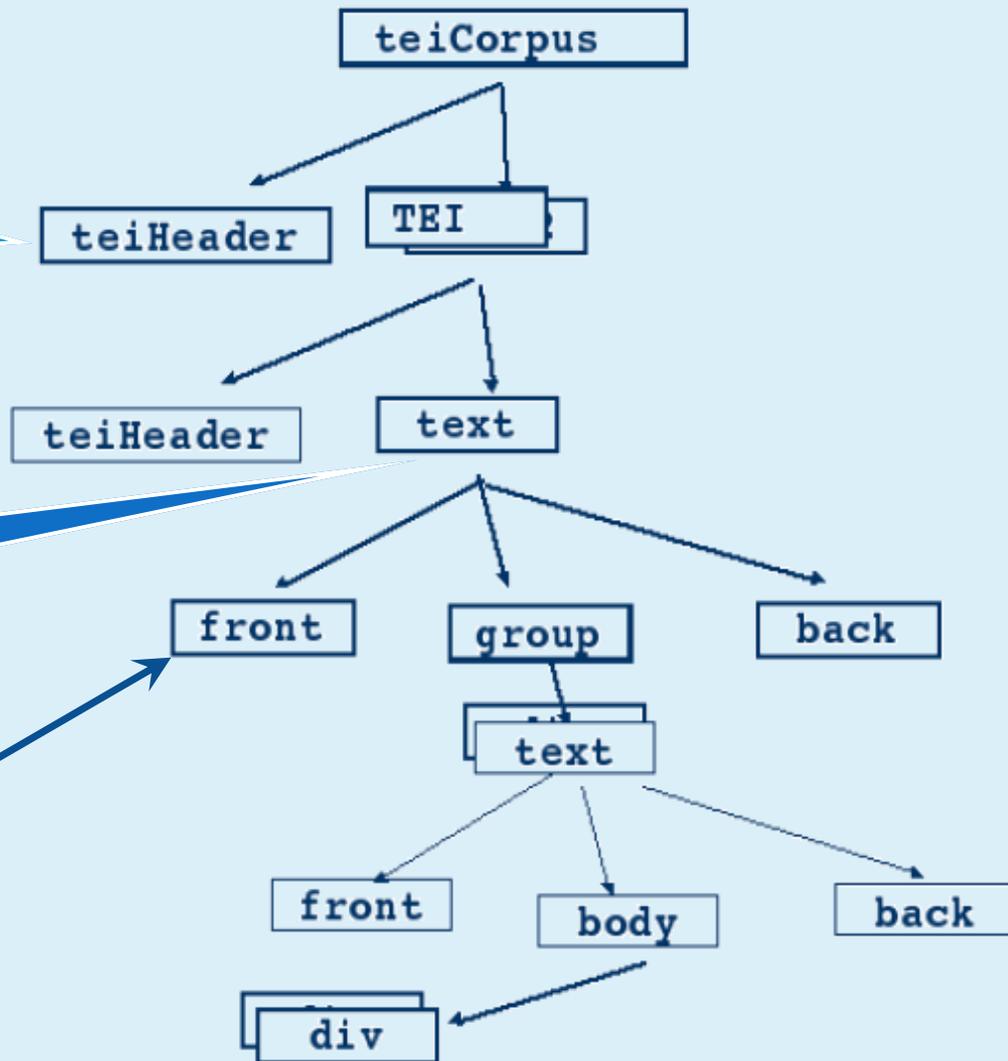
“back”

TEI : anatomie textuelle

Zone en-tête (métadonnées) :
Données bibliographiques, Techniques,
administratives, métadonnées sur la
ressources numériques ou analogique,

Le corps du document (texte, image,
son, vidéo). Subdivisé en:

Liminaires : page de titre, table des
matières, préface, dédicace etc



Annotation d'un poème (TEIVerse)

- ❑ Identifier dans un poème
 - *La mesure des vers*
 - *Les différents types de vers*
 - *Les groupes de vers (couplets, quatrains)*
 - *La strophe*
 - *La forme de la strophe*
 - *La rime*
 - *L'enjambement*
 - *Le rejet et le contre-rejet*
 - ...

Thé La structure d'un poème

Heureux qui, comme Ulysse, a fait un beau voyage

Heureux qui, comme Ulysse, a fait un beau voyage, A
Ou comme cestuy-là qui conquit la toison, B
Et puis est retourné, plein d'usage et raison, B
Vivre entre ses parents le reste de son âge! A

4 vers → un quatrain

Quand / ~~rev~~errai-je, hélas / de mon petit village / A 12 syllabes → un alexandrin
Fumer la cheminée, et en quelle saison, B
Reverrai-je le clos de ma pauvre maison, B
Qui m'est une province, et beaucoup davantage? A

Plus me plaît le séjour qu'ont bâti mes dieux, C
Que des palais Romains le front audacieux, C
Plus que le marbre dur me plaît l'ardoise fine: D

3 vers → un tercet

Plus mon Loir gaulois, que le Tibre latin, E
Plus mon petit Liré, que le mont Palatin, E
Et plus que l'air marin la douceur angevine. D

Recueil : Les Regrets
Joachim DU BELLAY (1522-1560)

Le choix du niveau de granularité peut varier entre grands segments et éléments plus petits

Annotation d'une pièce de théâtre (TEIDrama)

□ Identifier dans une pièce de théâtre :

- *L'interprétation et la mise en scène*

```
<acte 1>ACTE PREMIER
      <Trait>————</trait>
      <scène 1>SCENE I
<didascal>
  <groupe de lignes n°1>
    <ligne><lieu>Elseneur</lieu> — Une plate-forme devant le château</ligne>
      <ligne>FRANCISCO montant la garde, BERNARDO vient à lui</ligne>
    </fin de groupe de lignes>
</didascal>.
<groupe de lignes n°2>
<Acteur 1>BERNARDO . — Qui va là?</acteur>
<Acteur 2>FRANCISCO . — Non, répondez vous-même. Arrêtez-vous et faites-vous reconnaître. </acteur>
<Acteur 1>BERNARDO . — Vive le roi ! </acteur>
<Acteur 2>FRANCISCO . — Bernardo ? </acteur>
<Acteur 1>BERNARDO . — En personne. </acteur>
<Acteur 2>FRANCISCO . — Vous venez très soigneusement à votre heure. </acteur>
<Acteur 1>BERNARDO . — Minuit vient de sonner : va regagner ton lit, Francisco. </acteur>
<Acteur 2>FRANCISCO . — Pour cette délivrance, mille grâce. Le froid est aigre, et j'ai le cœur aussi</acteur>
</fin de groupe de lignes>
...
</fin de scène 1>
...
</ fin de l'acte 1>
```

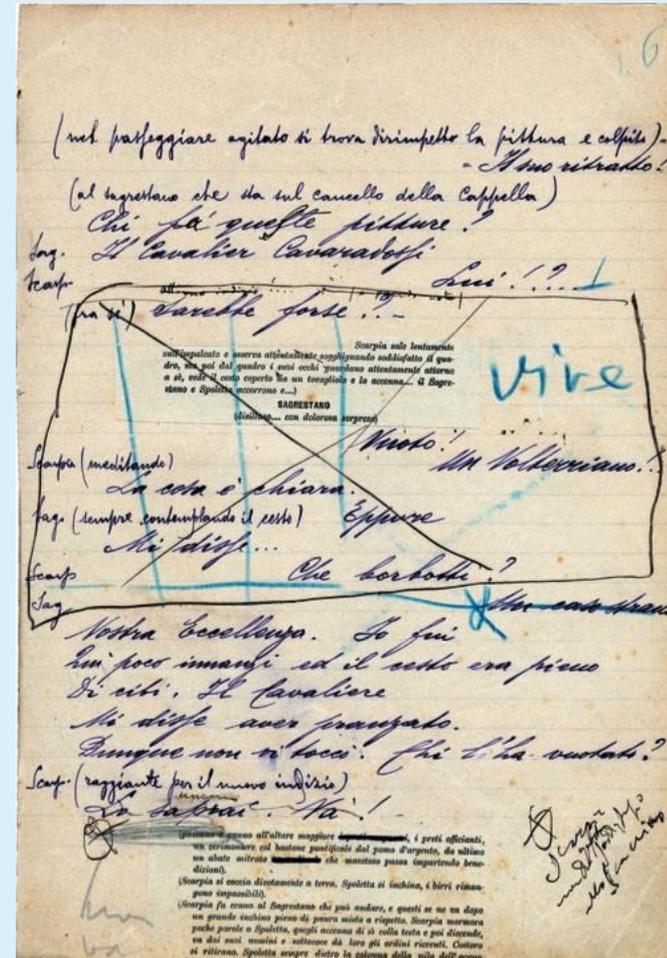
```
ACTE PREMIER
      SCÈNE I
Elseneur. — Une plate-forme devant le château.
FRANCISCO montant la garde, BERNARDO vient à lui.
BERNARDO. — Qui va là ?
FRANCISCO. — Non, répondez vous-même. Arrêtez-vous et faites-vous reconnaître.
BERNARDO. — Vive le roi !
FRANCISCO. — Bernardo ?
BERNARDO. — En personne.
FRANCISCO. — Vous venez très-soigneusement à votre heure.
BERNARDO. — Minuit vient de sonner : va regagner ton lit, Francisco.
FRANCISCO. — Pour cette délivrance, mille grâce. Le froid est aigre, et j'ai le cœur saisi.
```

Annotation d'un manuscrit (TEIManuscript)

- ❑ <surface> : une page, une stèle, tout objet avec une inscription
 - La surface contient des zones et des lignes
 - Elle a des coordonnées

- ❑ <zone>: Une aire de la superficie définie de façon arbitraire à des fins éditoriales. Les zones peuvent se superposer : la superposition est définie selon des coordonnées spatiaux
 - Peut contenir des <line>
 - Dispose de coordonnées

- ❑ <line> : une suite de texte identifiée de façon claire par l'éditeur
 - Peu contenir du texte et des <zone>
 - Ne dispose pas de coordonnées



Annotation d'un manuscrit (TEIManuscript)

Samarria is a Greek
island of water that
comes from the natural
springs of stilos, in
Crete.

1 April 2009

Feet birds in the park today.
Might write an article about
the Thick-billed Warbler.

Annotation d'un manuscrit (TEIManuscript)

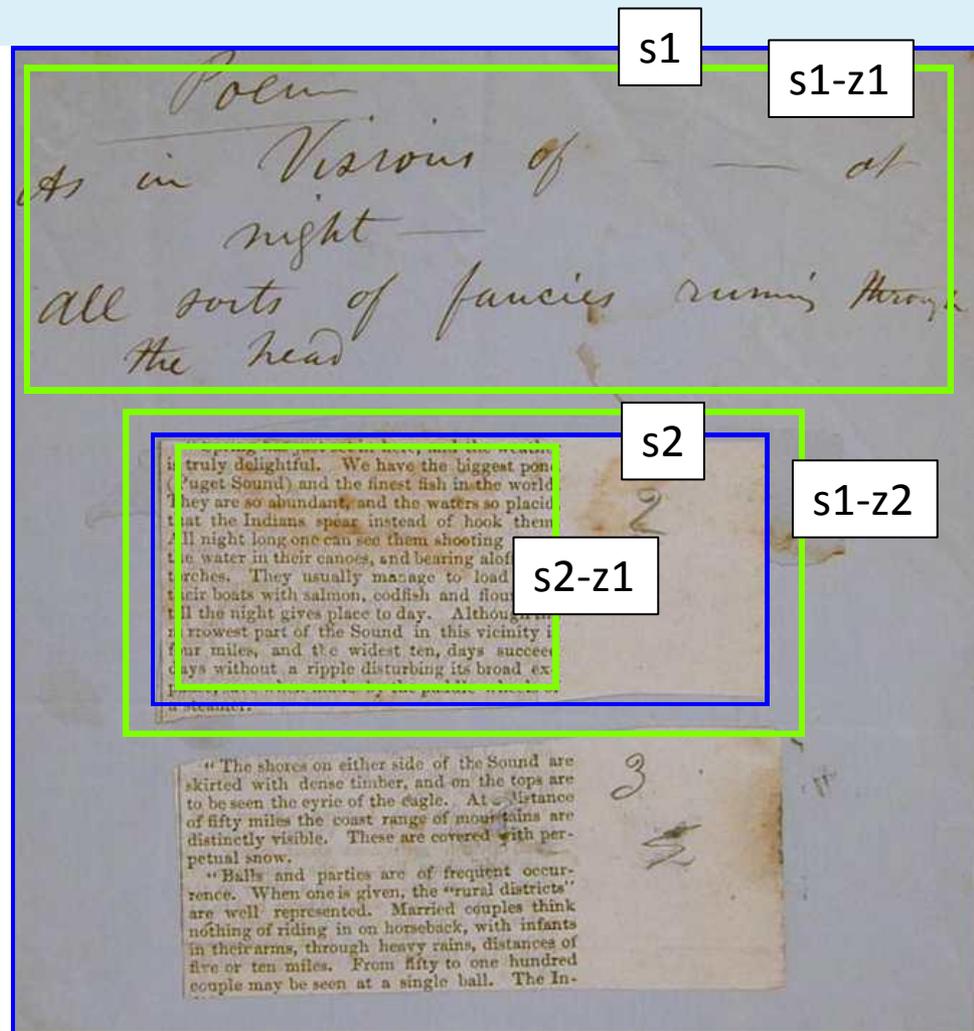
```
<surface
  ulx="0"
  uly="0"
  lrx="700"
  lry="1000">
  <!-- ... -->
</surface>
```

```
<zone
  ulx="93"
  uly="681"
  lrx="967"
  lry="1568">
  <graphic url="gb.jpg"/>
</zone>
```

```
<sourceDoc>
  <surface ulx="0" uly="0" lrx="200" lry="300">
    <zone ulx="10" uly="43" lrx="185" lry="84"
      rotate="0">
      <zone>
        <line rend="right"> 1 April 2009</line>
      </zone>
      <line>Fed Birds in the park today.</line>
      <line>Might write an article about</line>
      <line>the Thick-billed Warbler.</line>
    </zone>
    <zone ulx="9" uly="20" lrx="70" lry="60" rotate="90">
      <line>Samaria is a Greek</line>
      <line>brand of water that</line>
      <line>comes from the natural</line>
      <line>springs of Stilos, in</line>
      <line>Crete</line>
    </zone>
  </surface>
</sourceDoc>
```

Annotation d'un manuscrit (TEIManuscript)

```
<surface xml:id="s1" ulx="0" uly="0" lrx="50"
lry="50">
<zone xml:id="s1-z1" ulx="1" uly="1" lrx="10"
lry="10">
<line>Poem</line>
<!-- ... -->
<line>the head</line>
</zone>
<zone xml:id="s1-z2" ulx="4" uly="4" lrx="20"
lry="20">
<surface xml:id="s2" ulx="0" uly="0"
lrx="100" lry="100">
<zone xml:id="s2-z1" ulx="10" uly="10"
lrx="90" lry="95"> Spring has just set in here,
and the weather [...] a steamer </zone>
</surface>
</zone>
</surface>
```



Annotation d'un graphique (Graph)

Image Markup Tool

The screenshot displays the Image Markup Tool interface. On the left, a file list shows various image files (folio07r.jpg to folio25r.jpg). The main window shows a document page with handwritten text and a circular stamp. Annotations are made using red boxes and lines. A context menu is open over the annotations, showing options like 'Add zone', 'Link zone', and 'Delete zone'. A 'Dialog' window is open, showing XML code for indexing terms and a table of attributes.

XML Code:

```
57 <index
  xml:id="idx_indians"><term>I
  ndians</term><index><term>an
  d
  alcohol</term></index></inde
  x>
58 <index
  xml:id="idx_alcohol"><term>A
  lcohol</term><index><term>an
  d
  Indians</term></index></inde
  x>
59 <index
  xml:id="idx_proclamation"><t
  erm>Proclamation</term><inde
  x><term>against unauthorized
  sale of
  land</term></index></inde
  x>
60 Indians of Fraser's River;
  and also a Proclamation
  issued at
61 Fort Yale on the 15<hi
  rend="vertical-align: super;
  font-size: 80%;">th</hi> day
  of the same
62 month
63 <note xml:id="B00102">
```

Table:

Name	xml:id	Attribu	Value
#comment			&amp;...
#comment			&amp;...
#comment			..cm =====
div			type...
+ head			
+ opener			
- p	para_001	xml:...	
#text			I have the hono...
+ index			
#text			a Proclamation ...
+ hi		rend...	
#text			day of Septem...
+ note	B00101	n="...	
#text			prohibiting the ...
+ index	idx_indians	xml:...	
+ index	idx_alcohol	xml:...	
- index	idx_proclam...	xml:...	
- term			
...			Proclamation
+ index			
#text			Indians of Fras...
+ hi		rend	

Delete the selected zone

Annotation d'un graphique (Graph)

❑ Image Markup Tool

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>The Image Markup Logo</title>
    </titleStmt>
    <publicationStmt>
      <p></p>
    </publicationStmt>
    <sourceDesc>
      <p>377 x 259</p>
    </sourceDesc>
  </fileDesc>
  <encodingDesc>
  </encodingDesc>
</teiHeader>
```



Annotation MEI de notes de musique

```

<tuplet xml:id="t1" num="3" numbase="2">
  <beam xml:id="b1">
    <note xml:id="n1" pname="d" oct="5" dur="8" />
    <note xml:id="n2" pname="e" oct="5" dur="16" dots="1"/>
    <note xml:id="n3" pname="d" oct="5" dur="32" />
    <note xml:id="n4" pname="c" oct="5" dur="8" accid="s"/>
  </beam>
</tuplet>
<beam xml:id="b2">
  <tuplet xml:id="t2" num="3" numbase="2">
    <note xml:id="n5" pname="d" oct="5" dur="8" />
    <note xml:id="n6" pname="e" oct="5" dur="16" dots="1"/>
    <note xml:id="n7" pname="f" oct="5" dur="32" accid="s"/>
    <note xml:id="n8" pname="e" oct="5" dur="8" />
  </tuplet>
</beam>
  
```

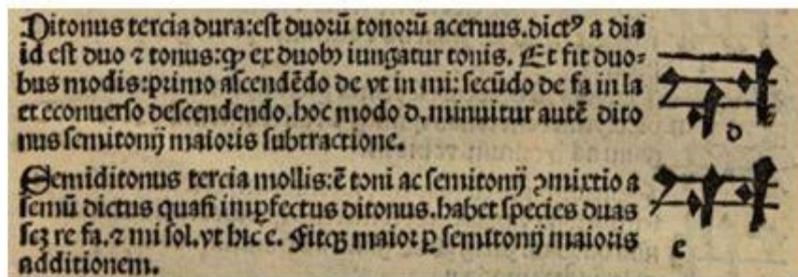


```

<g class="tuplet" id="t1" >
  <g class="beam" id="b1" >
    <g class="note" id="n1" >...</g>
    <g class="note" id="n2" >...</g>
    <g class="note" id="n3" >...</g>
    <g class="note" id="n4" >...</g>
  </g>
</g>
<g class="beam" id="b2" >
  <g class="tuplet" id="t2" >
    <g class="note" id="n5" >...</g>
    <g class="note" id="n6" >...</g>
    <g class="note" id="n7" >...</g>
    <g class="note" id="n8" >...</g>
  </g>
</g>
  
```

Application de la MEI au corpus

Intégration de la MEI à la TEI ?



Cochlaeus (1507), f. B1 r., facsimile.

Ditonus tercia dura, est duorum tonorum acervus, dictus a dia id est duo et tonus, que ex duobus jungatur tonis. Et fit duobus modis, primo ascendendo de ut in mi, secundo de fa in la et e converso descendendo, hoc modo d. Minuitur autem ditonus semitonii majoris subtractione.



Cochlaeus (1507), f. B1 r., édition TMG (2015).

```

<p>
<notatedMusic>
<mei:score>
  <mei:scoreDef key.sig="0">
    <mei:staffGrp>
      <mei:staffDef
        n="1" xml:id="P1"
        lines="3"
        clef.shape="F"/>
      </mei:staffGrp>
    </mei:scoreDef>
    <mei:section>
      <mei:staff n="1">
        <mei:layer n="1">
          <mei:note xml:id="d1e130" pname="c" oct="3"/>
          <mei:note xml:id="d1e142" pname="e" oct="3"/>
          <mei:note xml:id="d1e154" pname="f" oct="3"/>
          <mei:note xml:id="d1e166" pname="a" oct="3"/>
        </mei:layer>
      </mei:staff>
    </mei:section>
  </mei:score>
</notatedMusic>
Ditonus tercia dura [...]
</p>
  
```

Alternative II: Projeter la structure de la MEI à la TEI.

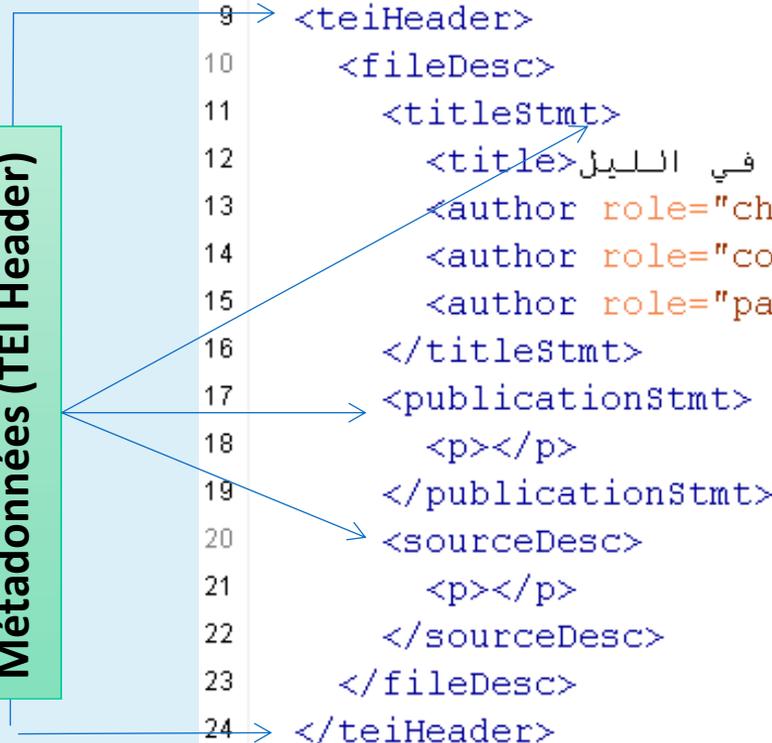
Annotation MEI de notes/texte de musique

Déclaration

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/sc
3 <?xml-stylesheet type="text/css" href="7-Taht-El-Yasmina.css"?>
4 <TEI xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
5   xmlns="http://www.tei-c.org/ns/1.0"
6   xmlns:math="http://www.w3.org/1998/Math/MathML"
7   xmlns:xi="http://www.w3.org/2001/XInclude"
8   xmlns:svg="http://www.w3.org/2000/svg">
9 <teiHeader>
10   <fileDesc>
11     <titleStmt>
12       <title>تحت اليا سمينة في الليل</title>
13       <author role="chanteur">الهادي الجويني</author>
14       <author role="compositeur">الهادي الجويني</author>
15       <author role="parolier">السيدة عزة</author>
16     </titleStmt>
17     <publicationStmt>
18       <p></p>
19     </publicationStmt>
20     <sourceDesc>
21       <p></p>
22     </sourceDesc>
23   </fileDesc>
24 </teiHeader>
  
```

Métadonnées (TEI Header)



Annotation MEI de texte de chanson

- تحت الياسمينه في الليل
- نسمة والورد محاذيني
- الأغصان عليا تـمـمـيل
- تمسحلي في دمة عيني
- تحت الياسمينه اتكيت
- عدلت العود وغنيت
- وتناظر دمعي وبكيت
- تفكرتك كيف كنت تجيني
- جنيته مزينها النوار
- فاحت من ريحت الأزهار
- تفكرتك شعلت النار
- عملت لهليبه في قلبي
- متوحش وحدي محتار
- لا قمره ولا حس أطيـار
- كان النسمة ع الأشجار
- توانس فيا وتواسيني

Stanza (Quatrin)

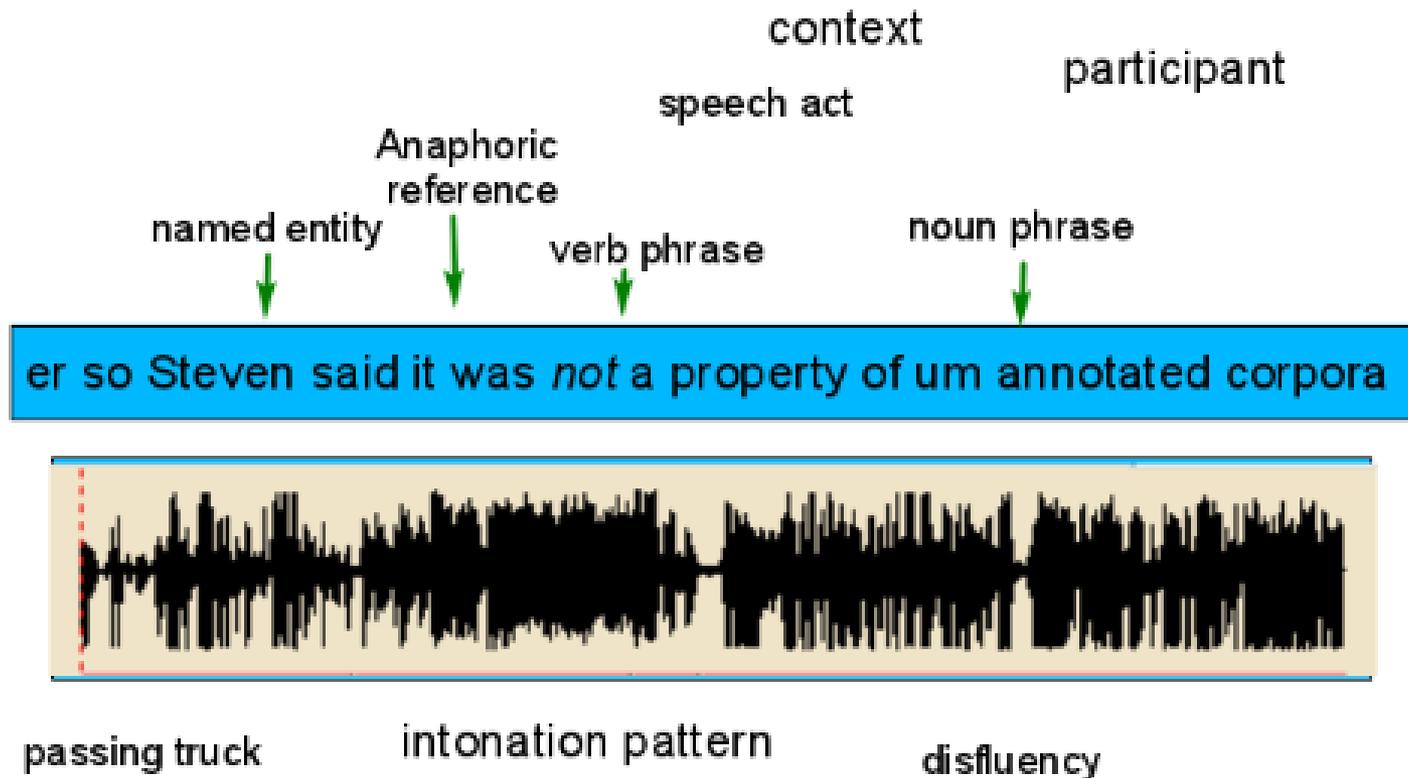
```

25 <text>
26 <body>
27 <head>تحت الياسمينه</head>
28 <lg type="stanza">
29 <1>تحت الياسمينه في الليل</1>
30 <1>نسمة والورد محاذيني</1>
31 <1>الأغصان عليا تـمـمـيل</1>
32 <1>تمسحلي في دمة عيني</1>
33 </lg>
34 <lg type="stanza">
35 <1>تحت الياسمينه اتكيت</1>
36 <1>عدلت العود وغنيت</1>
37 <1>وتناظر دمعي وبكيت</1>
38 <1>تفكرتك كيف كنت تجيني</1>
39 </lg>
40 <lg type="stanza">
41 <1>جنيته مزينها النوار</1>
42 <1>فاحت من ريحت الأزهار</1>
43 <1>تفكرتك شعلت النار</1>
44 <1>عملت لهليبه في كنيبي</1>
45 </lg>
46 <lg type="stanza">
47 <1>متوحش وحدي محتار</1>
48 <1>لاقمره ولا حس أطيـار</1>
49 <1>كان النسمة ع الأشجار</1>
50 <1>توانس فيا وتواسيني</1>
51 </lg>
52 </body>

```

Annotation d'un fichier son (TEISpeech)

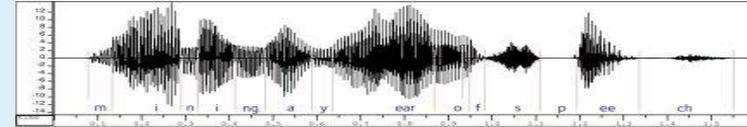
Un cas extrême -- il n y a pas de texte à annoter; on le crée en annotant.



Annotation d'un fichier son (TEISpeech)

- ❑ Un énoncé se définit comme une « séquence attribuée à un locuteur à un instant » : flux temporel « Timeline »

- ❑ Description très fine des différents phénomènes de communication oraux et non-oraux qui font partie du discours :
 - *pauses, chevauchements de paroles, changements d'intonation, de voix ou de langue, expressions vocalisées (tousser, rire, se moucher, grogner...), gestes, etc.*



```
<u who="#locuteur" sync="#T234">
  <seg type="interrupted">
    <kinesic>
      <desc>coughs</desc>
    </kinesic>
    <w>you</w>
    <w>must</w>
  </seg>
  <seg type="declarative">
    <w>you</w>
    <w>should</w>
    <w>let</w>
    <pause dur="short"/>
    <w>it</w>
    <w>be</w>
  </seg>
  <seg type="emphatic">
    <vocal>
      <desc>laughs</desc>
    </vocal>
    <w>please</w>
  </seg>
</u>
```

Annotation d'un fichier son (TEISpeech)

EXMARaLDA, par exemple

EXMARaLDA Partitur-Editor 1.5.1 [S:\TP-72\Publikationen\TEI_2010\Beispiel_EXMARaLDA.exb]

File Edit View Transcription Tier Event Timeline Format SFB 538/532 Help

très bien

00:00 00:01 00:02 00:03 00:04 00:05 00:06

	0 [00:00]	1 [00:01]	2 [00:02]	3 [00:03]	4 [00:04]	5 [00:05]	6 [00:06]
DS [sup]		Sister					
DS [v]	Oùay.	Très bien, très bien.			Ah oui?		
DS [en]	Oùay.	Very good, very good.					
DS [m]			right hand voice				
FB [v]			Alors ça dépend ((cough)) un petit peu				
FB [en]			That depends, then, a little bit				
FB [pho]				[ʁiˈpa]			

Done.

[13:30:37] Transcription 5:\TP-72\Publikationen\TEI_2010\Beispiel_EXMARaLDA.exb saved

EXMARaLDA: "Extensible Markup Language for Discourse



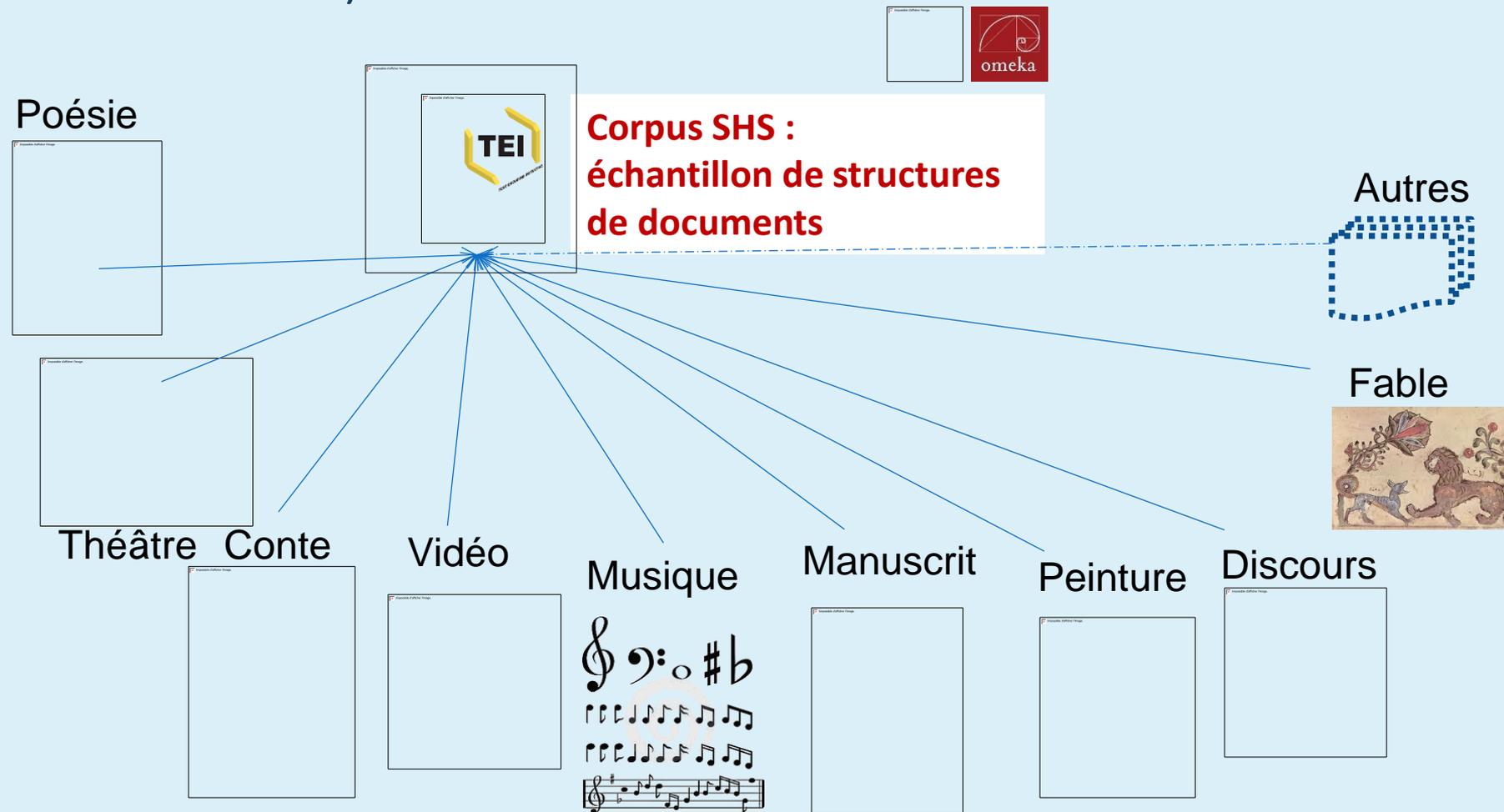
Annotation" <http://www.exmaralda.org/>

TEI & travail interdisciplinaire

- Un corpus TEI est une « entreprise » interdisciplinaire par excellence :
 - *Permet une grande liberté intellectuelle dans les choix d'encodage des corpus ;*
 - *Permet le travail collaboratif entre acteurs de plusieurs domaines ;*
 - *Permet l'élaboration de schémas de structuration transversaux à l'étude de plusieurs corpus.*

Modèle de collaboration interdisciplinaire

- ❑ Des schémas TEI par catégorie de ressource (eg. genres littéraires)



Modèle de collaboration interdisciplinaire

Spécialiste Arts & SHS



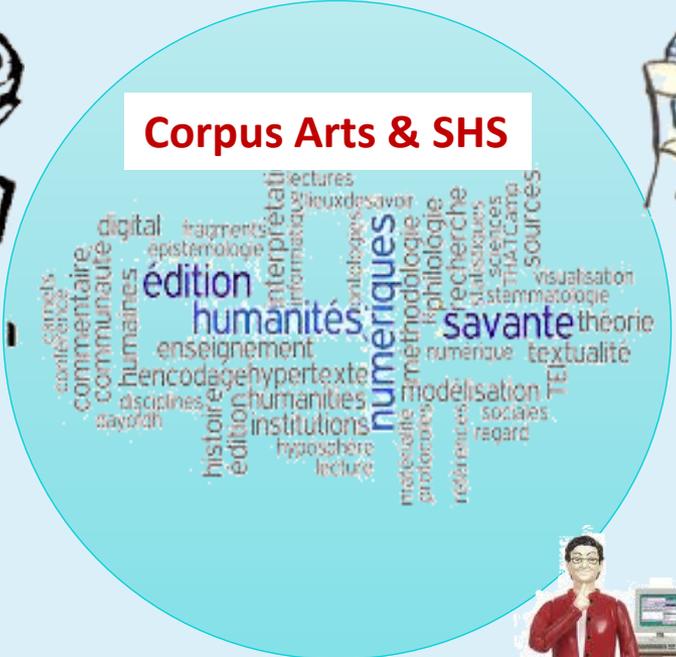
humnties www.fotosearch.com

Technicien XML/TEI



```
40 <sourceDoc>
41   <?Example TEI documents.</p>
42   <div>
43     <sourceDoc> bib1
44     </sourceDoc> bib1Full
45     <encodingDesc>
46       <editorialDecl>
47         <?Paragraph, page divisions and pu
48         Unless otherwise indicated by
49         in italics, and all SOCIALLED el
50       </editorialDecl>
51     </encodingDesc>
52     <crefDecl>
53     </div>
54   </sourceDoc>
```

Corpus Arts & SHS



Report
Titre-rapport
Résumé
Paragraphe
Chapitre
Titre-chapitre
Section
Titre-section
Paragraphe
Paragraphe

Spécialiste de l'Info-Com.



1
Définition/usage d'un schéma XML/TEI conforme a chaque type de corpus

3
Saisie des données des différents segments linguistique / sémantique du corpus

4
Génération du code XML/TEI des éléments du corpus

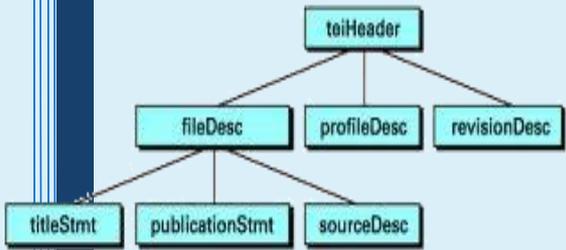


Schéma XML/TEI



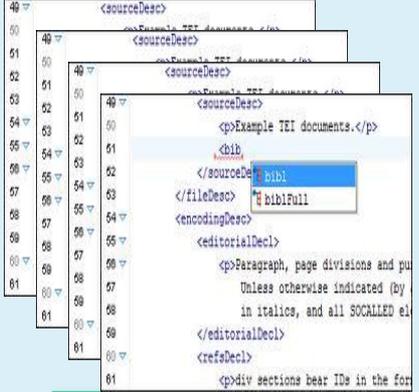
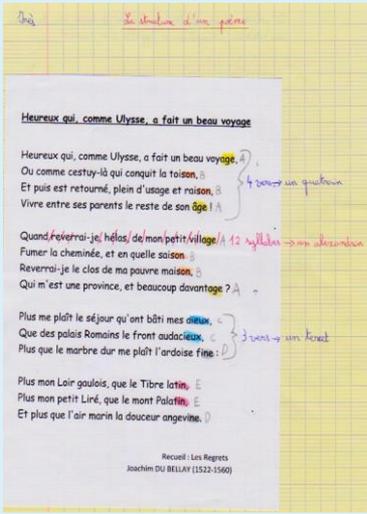
2
Production d'une interface graphique conforme au schéma XML/TEI



Briques sémantiques

I
N
T
E
R
F
A
C
E

G
R
A
P
H
I
Q
U
E



Encodage XML/TEI

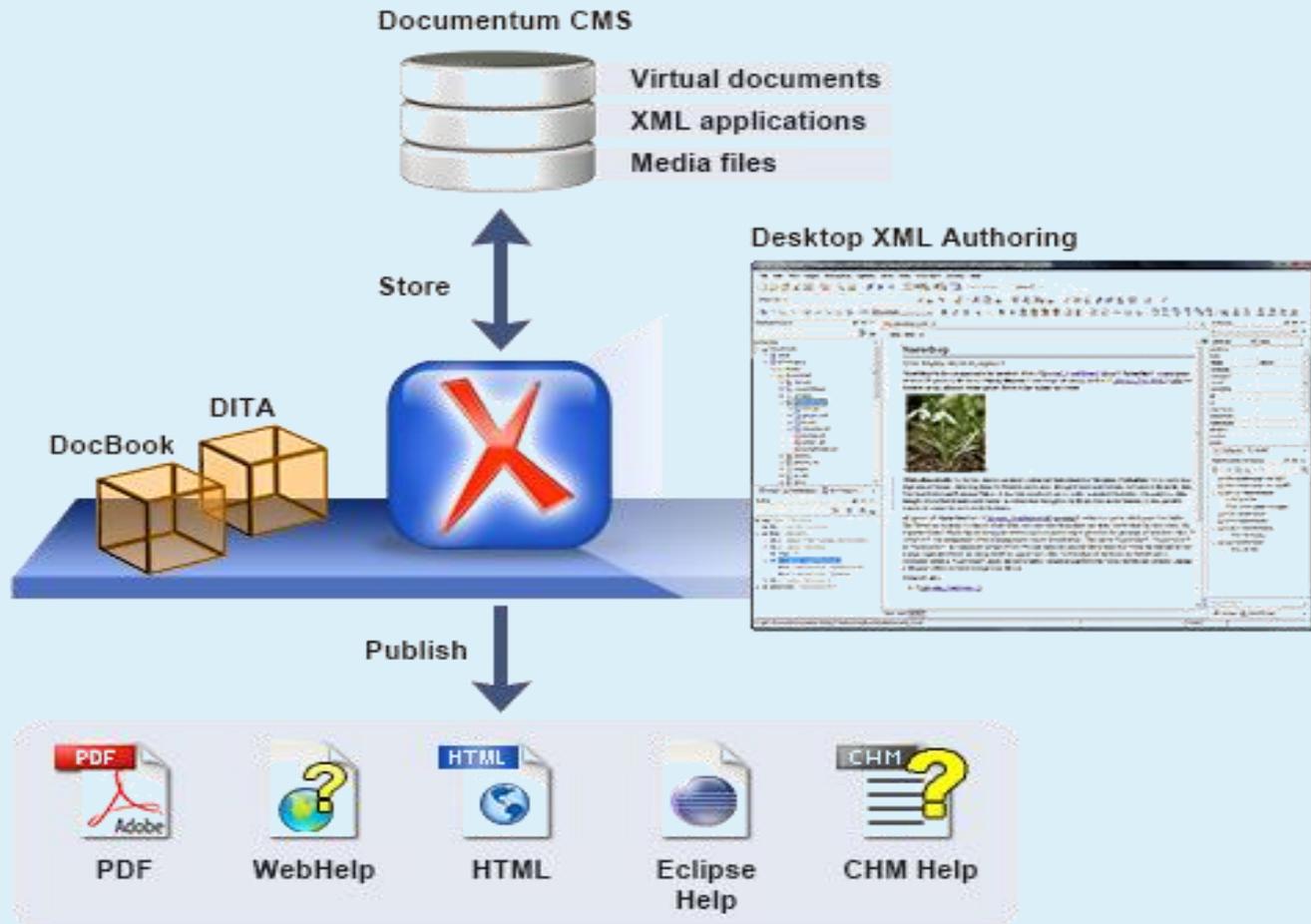


Usages multifonctions

- Réédition
- Corpus patrimoniaux
- Recherche

Des outils adaptés

- ❑ Dispositif CMS autour d'Oxygen : logiciel de production de corpus numérique en TEI ;



Des outils adaptés

- ❑ OXGARAGE : pour créer des schémas de documents TEI adaptés, interopérables et valides ;



TEI Roma: generating validators for the TEI

Set your parameters

[New](#) [Customize](#) [Language](#) [Modules](#) [Add Elements](#) [Change Classes](#) [Schema](#) [Documentation](#) [Save Customization](#)

Set your parameters

Title

Filename

Prefix for TEI pattern names in schema

Language English
 Deutsch
 Français
 Russian
 Svenska
 日本語
 中文

Author name

Description

Des outils adaptés

- ❑ OxGarage : pour convertir des formats de documents numériques de et vers la TEI ;

OxGarage Conversion

Select the format into which you want to convert your document

Convert from: ?



Documents

- Compiled TEI ODD Document
- DocBook Document
- Microsoft Word Document (.doc)
- Microsoft Word Document (.docx)
- ODD Document
- Open Office Text Document (.odt)
- OpenOffice 1.0 Text Document (.sxw)
- Plain Text (.txt)
- Rich Text Format (.rtf)
- TEI P4 XML Document
- TEI P5 XML Document
- TEI Tite XML Document
- WordPerfect Document (.wpd)
- xHTML Document

Convert to: ?

- Comma-Separated Values (.csv)
- ePub Document
- LaTeX Document
- Microsoft Excel Document (.xls)
- Microsoft Word Document (.docx)
- National Library of Medicine (NLM) DTD 3.0
- Open Office Spreadsheet (.ods)
- Open Office Text Document (.odt)
- OpenOffice 1.0 Spreadsheet (.sxc)
- OpenOffice 1.0 Text Document (.sxw)
- PDF Document
- Plain Text (.txt)
- RDF XML
- Rich Text Format (.rtf)
- Tab-Separated Values (.tsv)
- TEI P5 XML Document

CES (Corpus Encoding Standard)

- Un sous-ensemble de la TEI (DTD : Data Table Description) pour l'encodage des corpus :
 - *Un niveau d'encodage minimal que les corpus doivent atteindre pour être considérés comme standardisés ;*
 - *Des conventions pour un encodage plus étendu pour l'annotation linguistique ;*
 - *Une architecture générale pour représenter les corpus avec des annotations linguistiques*

- La DTD CES (schema) complète la TEI pour :
 - *l'étiquetage grammatical (CESAna) :*
 - *l'alignement de corpus (CESAlign) :*

ISO 24624:2016

□ ISO 24624:2016 - Gestion des ressources linguistiques -- Transcription du langage parlé :

- *énonce des règles de représentation des transcriptions d'enregistrements audio et vidéo d'interactions parlées ;*
 - transcription pour des études sociolinguistiques, l'analyse de conversation, la dialectologie, la linguistique de corpus, la lexicographie de corpus, les technologies langagières, les études qualitatives en sciences sociales, et aux autres données de transcription d'enregistrements du langage parlé ;
- *rattache les données transcrites à des normes de corpus annotés ;*
- *ne s'applique pas aux autres formes de transcription et surtout pas aux transcriptions de manuscrits ;*

Un exemple d'initiative

- ❑ Exemple de l'Initiative de la COMUE de Lyon (projet préparé l'USR 3439 MOM, l'UMR 5189 HiSoMA et l'UMR 5648 CIHAM)
 - *Programme d'acculturation digitale :*
 - Mettre en synergie des différents métiers dans le HN ;
 - Faire évoluer les cloisonnements disciplinaires et faciliter la construction des dimensions transdisciplinaires et innovantes des projets en HN,
 - Concevoir des outils en phase avec les exigences épistémologiques des recherches en HN ;
 - Rendre transparents les présupposés théoriques sous-jacents à ces outils, qui ne sont jamais purement techniques mais intègrent nécessairement des choix intellectuels que leurs utilisateurs doivent pouvoir maîtriser ;
 - *Construire un écosystème compétitif en HN :*
 - Produire des données partageables et réutilisables (algorithmes [API]) ;
 - Permettre la réutilisabilité numérique grâce aux normes internationales (TEI, OAIS)

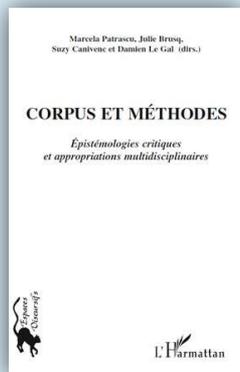
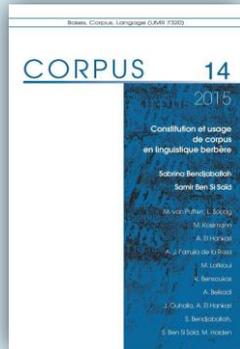


Un exemple d'initiative

- ❑ Exemple de l'Initiative de la COMUE de Lyon (projet préparé l'USR 3439 MOM, l'UMR 5189 HiSoMA et l'UMR 5648 CIHAM)
 - *Un incubateur de recherche pour les humanités numériques*
 - personnels spécialisés, aide à la gestion de projets numériques ;
 - formation aux bonnes pratiques et aux standards de description ;
 - mise en œuvre des procédés d'encodage (TEI) ;
 - anticipation des formats et procédures d'archivage des données ;
 - labellisation par les réseaux et infrastructures françaises et européennes (portails Isidore, Europeana, Darjah) ;
 - Alignement sur les grandes infrastructures de recherche européenne (Darjah) et française (TGIRHuma-num) fondées sur les meilleures technologies et standards ;
 - développement avéré de compétences reconnues à un niveau international dans le domaine du web sémantique et de l'open data.



Quelques références



- ❑ La revue CORPUS (openedition.org)
- ❑ Poudat C, Landragin F. Explorer un corpus textuel: Méthodes - pratiques - outils. Paris: De Boeck Supérieur; 2017. 243 p.
- ❑ Wigham CR, Ledegen G. Corpus de communication médiée par les réseaux: Construction, structuration, analyse. Editions L'Harmattan; 2017. 260 p.
- ❑ Comby É. Corpus de textes : composer, mesurer, interpréter. ENS Éditions; 2016. 194 p. x
- ❑ Patrascu M, Brusq J. Corpus et méthodes: épistémologies critiques et appropriations multidisciplinaires. Harmattan; 2011. 210 p.

