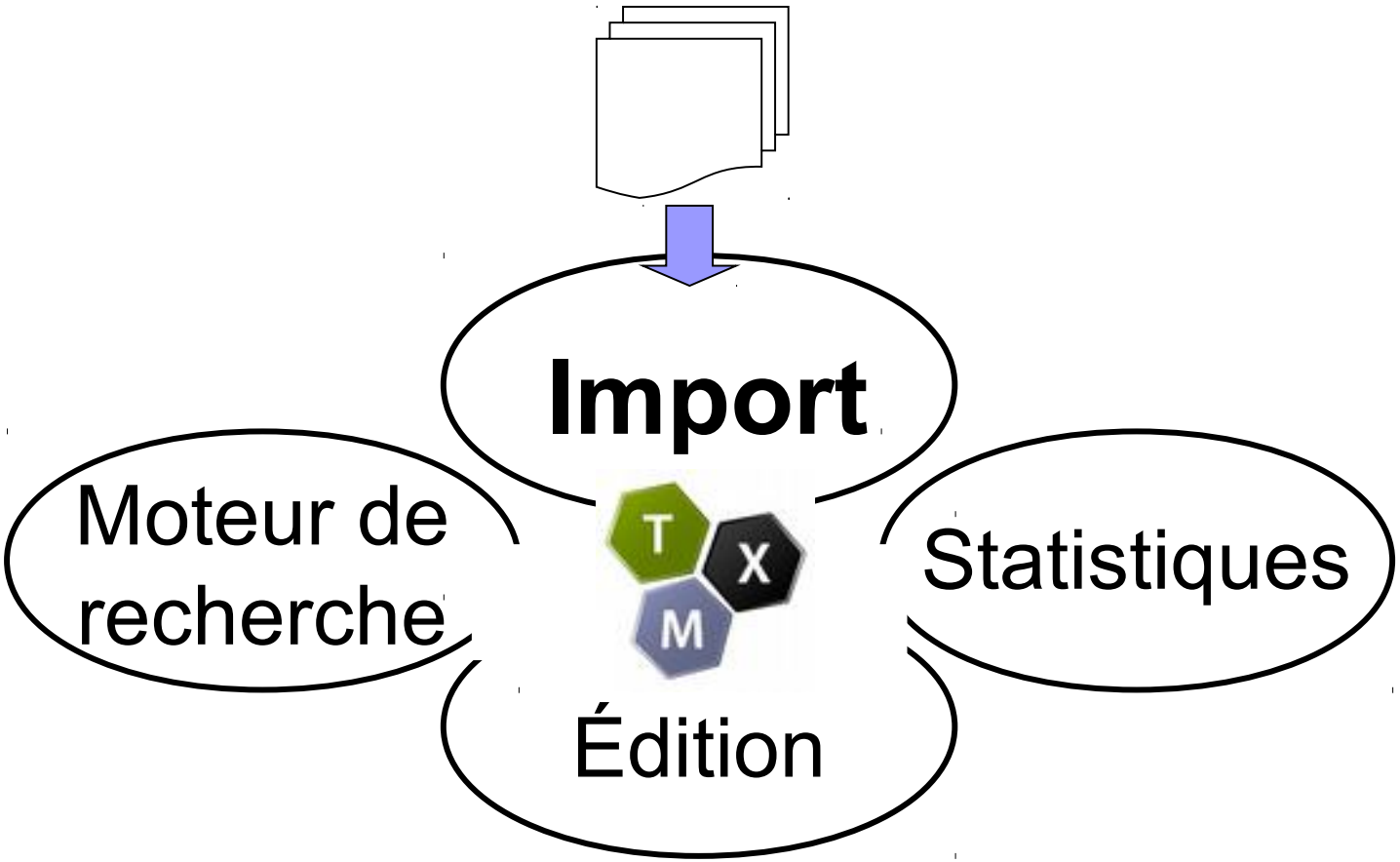
Atelier préparation de corpus et import dans TXM

Serge Heiden et Alexei Lavrentev
ENS de Lyon
slh@ens-lyon.fr

UMR5317 IHRIM, 2014-2017

Import dans TXM



3 familles de corpus

A) Corpus de **textes écrits** (TXT, XML, TEI) éditions alignées avec images de facsimilés

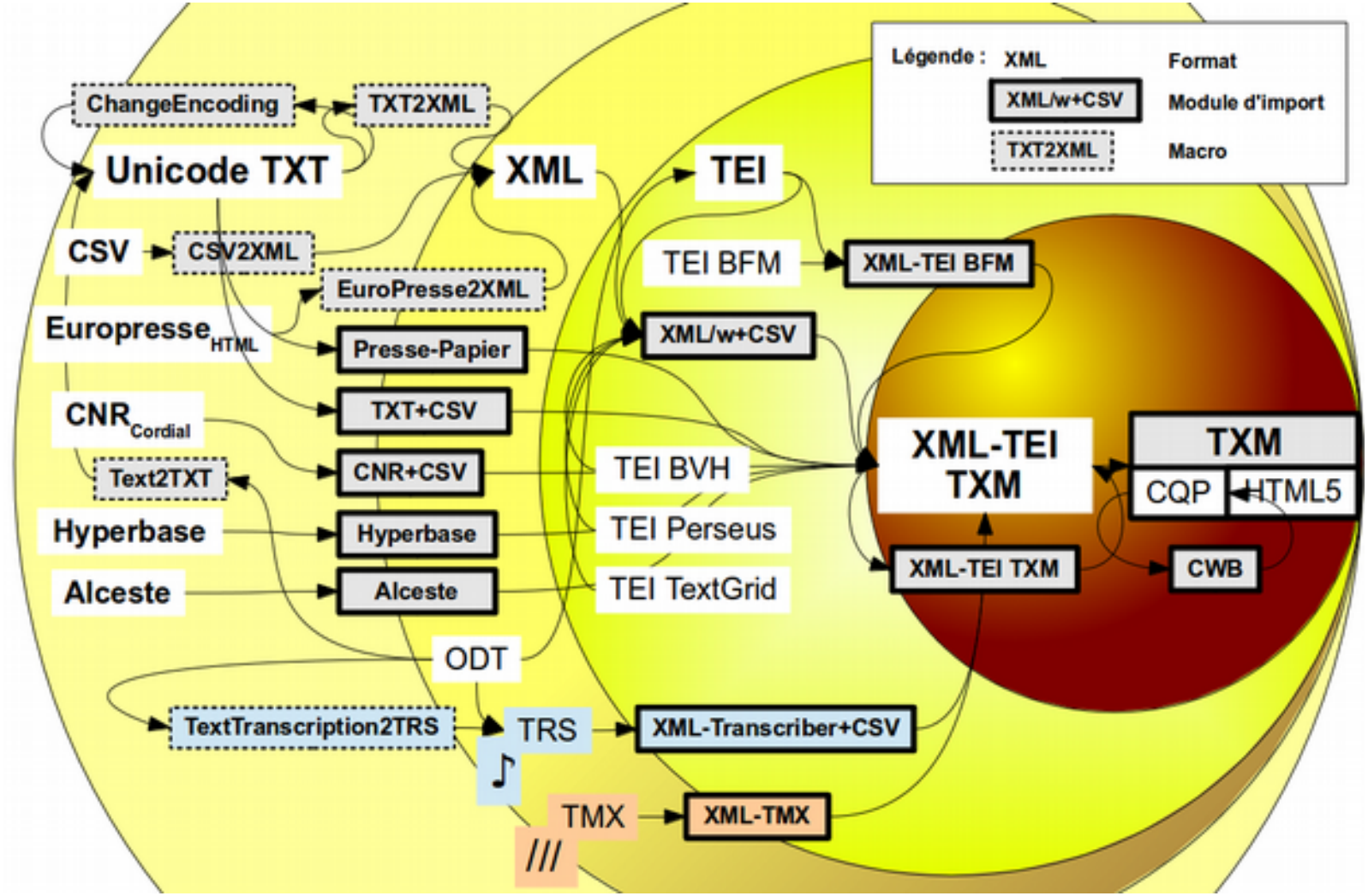
B) Corpus de **transcriptions d'enregistrements** (TRS), éventuellement *synchronisées* avec le son ou la vidéo

C) Corpus **multilingues alignés** (TMX), au niveau d'une structure textuelle comme la phrase ou le paragraphe

Modèle de corpus TXM

- **Unités textuelles** (livre, article, entretien…)
  Métadonnées (auteur, date, domaine, genre…)
  - **Structures internes** (phrase, paragraphe, sections…)
    Propriétés (numéro, titre...)
    - **Unités lexicales** (mots, mots composés)
      Propriétés (graphical form, lemma, part of speech…)
  - **Plans textuels**
    - Hors-texte (en-tête TEI, notes éditoriales)
    - Tours de parole, discours direct...
    - Langue principale (français…), Langue secondaire (latin…)
- Outils de TAL impliqués (lemmatiseurs…)
- Édition de texte pour le retour au texte
  - Pagination (sauts de pages)
  - Mise en page (styles), Média (Image, Audio, Vidéo)
- Alignement (corpus parallèles)

Environnement d'import TXM



TXM import modules

corpus input formats

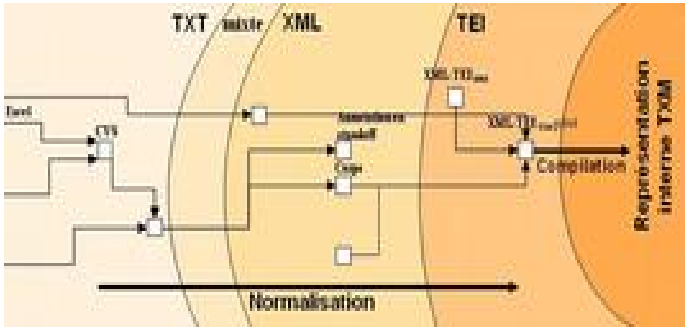
- Formats propriétaires divers : Hyperbase, Alceste, CNR (Cordial)
- *Calibre* – ePub

- Copier/Coller
- TXT Unicode+CSV (metadata) : dossier de fichiers texte brut
- XML/w+CSV : dossier de fichiers XML
- (TXM 0.7.8) **XML-XTZ** : reconfiguration des textes, éditions stylées

- XML-TEI P5 **BFM** : personnalisation de la TEI
- XML-TEI P5 **FRANTEXT (textes)**
- XML-TEI P5 **FRANTEXT (résultats de requêtes)**
- XML-TEI-TXM : XML compatible TEI+TAL (pivot)

- XML-Transcriber+CSV – transcriptions audio alignées
- *XML-TMX* – corpus multilingues alignés
- *XML-PPS*-Factiva – portail de presse

Basic import & analysis workflow



- TXT texts directory
- XML texts + metadata
- NLP Tagged texts
- XML-TXM TEI texts
- Contrasts : sub-corpus & partition
- Structures
- Lexical facets

TreeTagger

TXM

Modèle de corpus TXM

- **Unités textuelles** (livre, article…)
  Métadonnées (auteur, date, domaine, genre…)
  - **Structures internes** (phrase, paragraphe, sections…)
    Propriétés (numéro, titre...)
    - **Unités lexicales** (mots)
      Propriétés (graphical form, lemma, part of speech…)
  - **Plans textuels**
    - Hors-texte (entête TEI) / hors texte à éditer (notes)
    - Tours de parole, discours direct...
    - Langue principale (français…), Langue secondaire (latin…)
- Outils de TAL impliqués (lemmatiseurs…)
- Édition de texte pour le retour au texte
  - Pagination (sauts de pages)
  - Mise en page (styles), Média (Image, Audio, Vidéo)
- Alignement (corpus parallèles)

Carte des niveaux d'import TXM

	TXT	**XML/w**	**XTZ**
Unités Textuelles	fichiers	fichiers	fichiers / XSL split-merge
Métadonnées	CSV	CSV	CSV
Mots	brut	<w>?	<w>?
Structures	-	toutes	toutes
Plans	-	XSL frontale	interface / XSL frontale & post-tokenisation