Note sur le fonctionnement de l'annotation sémantique dans Transkribus

Tags prédéfinis dans Transkribus

Transkribus possède des tags prédifinis qui peuvent être utiles pour le projet :

Nom	Utilisation	Export TEI
abbrev	annoter une abréviation	abbrev devient toujours le groupe d'éléments `<choice><expan></expan><abbr>...</abbr></choice>` contenant la sélection annotée. Sa propriété "expansion" est utilisé pour indiquer un contenu dans `<expan>`. Si elle est n'est pas remplie, reste vide.
date	annoter une date	date devient un élément `<date>`. Ses propriétés "year", "month" et "day" sont toujours exportées vers les attributs @year, @month et @day, lorsque qu'elles possèdent des valeurs.
organization	annoter une organisation	organization devient l'élément `<orgName>`. Il ne possède pas de propriété.
place	annoter un lieu	place devient l'élément `<placeName>`. Sa propriété "country" devient le sous-élément `<country>`. En revanche, sa propriété "placeName" n'est pas exportée.
person	annoter une personne	person devient l'élément `<persName>`. La propriété "occupation" n'est pas exportée en TEI. Lorsque l'on ajoute des propriétés supplémentaires, celles-ci ne sont pas prises en charge lors de l'export. Ses autres propriétés sont exportés comme suit : "dateOfBirth" = `<birth>` ; "dateOfDeath" = `<death>` ; "firstname" = `<forename>` ; "lastname" = `<surname>` ; "notice" = `<notice>`
work	annoter une référence bibliographique	work devient l'élement `<work>` et ses propriété "creator", "year" et "title" sont respectivement exportées dans les attributs @creator, @year et @title.
unclear	annoter une incertitude sur la lecture du texte	unclear devient l'élément `<unclear>` et sa propriété "alternative" devient l'attribut @alternative.
commentaires*	ajouter un commentaire	les commentaires sont exportés au sein d'un élément `<note>` directement accolé à la sélection sur laquelle il porte.

* Les commentaires ne sont pas gérés via les tag mais par le biais d'une interface spécifique.

La gestion de la TEI par Transkribus n'est pas satisfaisante pour le moment car toutes les règles ne sont pas respectées. L'export des annotations réalisée avec les tags par défaut ne fonctionne pas parfaitement car certaines propriétés sont ignorées par Transkribus au moment de la transformation.
Les nouveaux tags ne sont pas vérifiés par Transkribus, et il n'est pas possible de paramétrer la manière dont ils sont exportés : les tags personnalisés ne sont donc pas conformes à la TEI.
Afin d'éviter les ambiguités d'usage pour les annotateur□rice□s, il semble préférable d'utiliser des tags nommés en français, plutôt que de nommer les tags par rapport à leur forme souhaitée en TEI au moment de l'export.

Dans la mesure où Transkribus permet de créer une liste de tags "à part", il semble utile de créer un jeu de tags spécifiques au projet Time Us afin de les faire parfaitement correspondre aux besoins de l'annotation.

Tags du projet Time Us

Seuls les tags **date**, **place** et **organization** sont repris des tags par défaut, car ils ne sont pas ambigus et leur export TEI fonctionne correctement.
Afin de différencier les Tags du projet Time Us des tags prédéfinis (ou d'autres tags installés par l'utilisateur), ces nouveaux tags sont précédés du préfixe "TU_".

Tag Transkribus	Export XML par Transkribus	Forme de balise souhaité
TU_incertitude	`<TU_incertitude>`	`<certainty cert='low/medium/high'>`
TU_adresse	`<TU_adresse>`	`<address>`
TU_document + @type	`<TU_document type=''>`	`<bibl type=''>`
TU_personne	`<TU_personne>`	`<persName>`
TU_heure	`<TU_heure>`	`<time>`
TU_montant + @type	`<TU_montant type=''>`	`<measure type='sum' subtype=''>`
TU_quantite + @unit	`<TU_quantite>`	`<measure type='count' unit=''>`
TU_occupation + @normal	`<TU_occupation normal=''>`	`<choice><orig><rs type="occupation">[occupation]</rs></orig><reg><rs type="occupation">[@normal]</rs></reg></choice>`
TU_statut	`<TU_statut>`	`<rs type='workerStatus'>`
TU_duree	`<TU_duree>`	`<rs type='duration'>`
TU_produit	`<TU_produit>`	`<rs type='product'>`
TU_statutMatrimonial	`<TU_statutMatrimonial>`	`<rs type='matStatus'>`
TU_tache	`<TU_tache>`	`<rs type='task'>`
TU_typeRemuneration	`<TU_typeRemuneration>`	`<rs type='revenue-type' subtype=''>`
TU_remuneration	`<TU_remuneration>`	`<s type='revenue'>`
commentaire	`<note>`	`<!-- -->`
organization	`<orgName>`	`<orgName>`
date	`<date>`	`<date>`
place	`<placeName>`	`<placeName>`

Explications pour chacun des éléments

Echappement : TU_incertitude

Afin d'obtenir une annotation sémantique de qualité, il est important de prévoir des mécanismes qui permettent aux annotateur□rice□s de signaler les passages problématiques. C'est le but de ce tag, qui permettra de rapidement faire remonter les points problématiques afin qu'ils soient traités par des annotateur□rice□s plus expertes et en mesure de trancher.

L'élément "certainty" est amplement suffisant pour signaler un passage problématique. Dans le cadre d'une signalement simple de passages difficiles, il ne semble pas utile de laisser à l'annotateur□rice le choix du niveau de certitude : on pourra par défaut attribuer la valeur "medium"

à l'attribut @cert. On peut en revanche utiliser cette échelle et modifier la valeur manuellement pour identifier les cas encore problématiques après un traitement par un expert.

Commentaire

L'annotateur□rice doit pouvoir commenter ses choix ou préciser la raison d'une incertitude. Le contenu de ce texte ne devrait pas apparaître dans le document. Il semble logique d'utiliser l'outil de Transkribus prévu à cet effet, mais l'élément "note" fait généralement partie du flux d'informations de premier niveau, c'est pourquoi le transformer en commentaire XML permettrait de le ramener au niveau du processus d'annotation.

Lieux : place

Les documents contiennent de nombreuses références à des noms de lieux.

Organisation : organization

Les documents contiennent de nombreuses références à des noms d'établissements, d'institutions ou d'organisations.

Adresses : TU_adresse

Les lieux et les établissement sont souvent associés à une adresse qu'il sera possible de leur rattacher ensuite **par le biais du TAL**.

Références externes : TU_document

Les documents contiennent des références à des documents externes comme des lois, des tarifs, des articles de presse. Ceux-ci peuvent consituer des supports pour mieux comprendre le calcul à effectuer pour recomposer une rémunération, ou bien pour identifier de nouvelles sources. C'est la raison pour laquelle leur annotation est utile. Les références externes peuvent être typées grâce à l'attribut @type afin de préciser la nature de cette référence. Seule la lecture des documents et l'analyse des situations permet d'établir une liste complète de types de références externes.
- presse - tarif - loi

Il s'agit généralement de références à des documents, c'est pourquoi l'élément "bibl" semble satisfaire. On pourrait faire le choix d'utiliser l'élément "rs" avec comme valeur de @type "externalRef" et la catagorie en @subtype.

Personne : TU_personne

Les documents contiennent de nombreuses mentions de personnes. L'annotation manuelle des personnes se concentre uniquement sur les noms de personnes : pas sur des situations de coréférence. L'élément "persName" suffit pour annoter cela. Les informations présentes à propos des personnes (leur statut matrimonial, leur genre, etc) pourront leur être reliés **par le biais du TAL**.

Statut Matrimonial : TU_statutMatrimonial

L'annotation des statuts matrimoniaux est surtout utile dans le cas des femmes. Elle doit permettre d'établir une liste de vocabulaire pour pouvoir identifier des expressions qui permettent de comprendre le statut d'une personne. L'élément "rs" est adapté pour ce genre d'annotation : il peut s'agir d'une expression à un ou plusieurs mots. Ce type d'information pourra être relié à la personne à laquelle le statut fait référence **par le biais du TAL**.

Heure, durée et date : TU_heure, TU_duree, date

Contrairement à l'expression de la durée, les heures sont des informations très explicites et dont la forme varie peu. Leur annotation permet de réaliser un calcul pour mieux comprendre l'organisation de la journée de travail en terme de temps. **Mais cette information n'est peut-être pas pertinente**...

Les informations de durée ou de rythme, en revanche, sont exprimés de manières très variées. Elle permettent de préciser le temps de travail associé à une rémunération, voire un rythme de travail attendu. Comme des informations prennent beaucoup de formes, leur annotation peut poser des problèmes. Il s'agit presque toujours cependant d'expressions relativement courtes et qui ne sont pas composées d'autres types d'informations, c'est pourquoi on utilise l'ément "rs" plutôt que l'élément "s".
Les dates sont des informations homogènes qui permettent de contextualiser les informations. Elles sont cependant souvent exprimées de manière partiellement implicites. On ne retient que les dates les plus précises possibles, celles qui font référence à une année, afin d'avoir des informations utiles pour la contextualisation.

Rémunération : TU_remuneration

L'annotation des passages évoquant des rémunérations doit permettre dans un premier temps de collecter des informations sur les différentes manières d'exprimer lesdites rémunérations. Il s'agit généralement de morceaux de phrase(s) plus ou moins longs et qui contiennent d'autres éléments annotés. La rémunération est une information qui émane de la mise en commun de tous ces éléments qui la composent, c'est pourquoi l'élément "s" est beaucoup plus approprié que l'élément "rs". Les informations sur les rémunérations seront recomposées et normalisées **par le biais du TAL**.

Montant et quantité : TU_montant, TU_quantite

Les montants et les quantités sont des annotations portant sur des nombres. Ce sont deux sujets d'annotation traduits par l'élément "measure" dont l'attribut @type est essentiel. La valeur de cet attribut sera "currency" dans le cas d'un montant d'argent, et "count" dans le cas où il s'agit uniquement d'exprimer une quantité (de personnes, de produit, etc). Les quantités sont associées à des unités qui varient beaucoup, c'est pourquoi l'attribut @unit doit être systématiquement utilisé. L'attribut @subtype permet de préciser si ce montant ou cette quantité est absolue ou relative. Sa valeur sera donc soit "absolute", soit "relative".

Type de rémunération : TU_typeRemuneration

L'annotation des informations sur les types de rémunération doit permettre de distinguer les différentes situations conduisant à une rémunération. Le type est généralement exprimé dans une expression courte, c'est pourquoi on utilise l'élément "rs". L'attribut @subtype sert à préciser le type de rémunération de manière normalisée. Seule la lecture des documents et l'analyse des situations permet d'établir une liste complète de types de rémunération : - solde - avance - total - indemnité_judiciaire - indeminité_syndicale - solidarité

Métiers : TU_occupation et TU_status

L'annotation manuelle des expressions désignant des métiers doit permettre de créer une liste de vocabulaire afin de réaliser une annotation automatique de ces métiers. Cette liste doit également permettre de générer des données statistiques sur la fréquence d'utilisation des noms de certains métiers, voire sur les régions où ces noms sont utilisés. Ces noms de métiers sont souvent genrés. Leur annotation doit donc également permettre d'identifier des informations sur la répartition des tâches et des métiers en fonction du genre. Les noms de métiers sont souvent composites, puisqu'on désigne un statut (ex : patron, ouvrier), puis un secteur d'activité (ex : tissage mécanique, tissage à bras, etc). Cela peut être une difficulté. L'attribut "normal" est essentiel afin de recomposer des formes normalisées de ces expressions que la structure de la phrase déforme. Sans cette normalisation, on perd des informations.

L'annotation des statuts de travailleurs, tels que "patron", "chef d'atelier", "apprenti", doit permettre de mieux comprendre la situation d'un□e travailleur□se, en particulier en rapport avec sa rémunération ou ses tâches.

Produits et tâches : TU_produit, TU_tache

Les produits et les tâches (actions ou verbes) sont généralement évoqués dans le contexte d'une rémunération, mais pas toujours. Les produits peuvent être associés à une quantité. Il s'agit d'établir des listes de vocabulaire pour mieux comprendre tous les types d'activité autour du textile, et permettre par exemple de rassembler des passages parlant des mêmes types de produits pour pouvoir les comparer. Ces informations pourront être croisés avec le temps de travail qu'ils induisent, ou la rémunération à laquelle ils donnent lieu par le biais du TAL.