Humanités numériques : structuration des données et des documents textuels

E. Rouquette

Cours 1 – 5 février 2025

Introduction

Présentation du cours

Ressources du cours : https://github.com/Enimie/TNAH1_XML

Présentation du cours

Ressources du cours : https://github.com/Enimie/TNAH1_XML

Objectifs du cours

- Comprendre ce qu'est le balisage d'un texte
- Savoir utiliser Markdown et transformer un fichier `.md` au format voulu (Pandoc) (syntaxe de base)
- Comprendre les enjeux liés aux XML
- Savoir lire et composer un document XML de base
- Connaître les technologies liées à XML
- Apprendre les éléments de base dans les langages TEI et EAD
- Première approche des méthodes permettant exploiter les documents XML (XSLT, XPath)

Présentation du cours

Ressources du cours : https://github.com/Enimie/TNAH1_XML

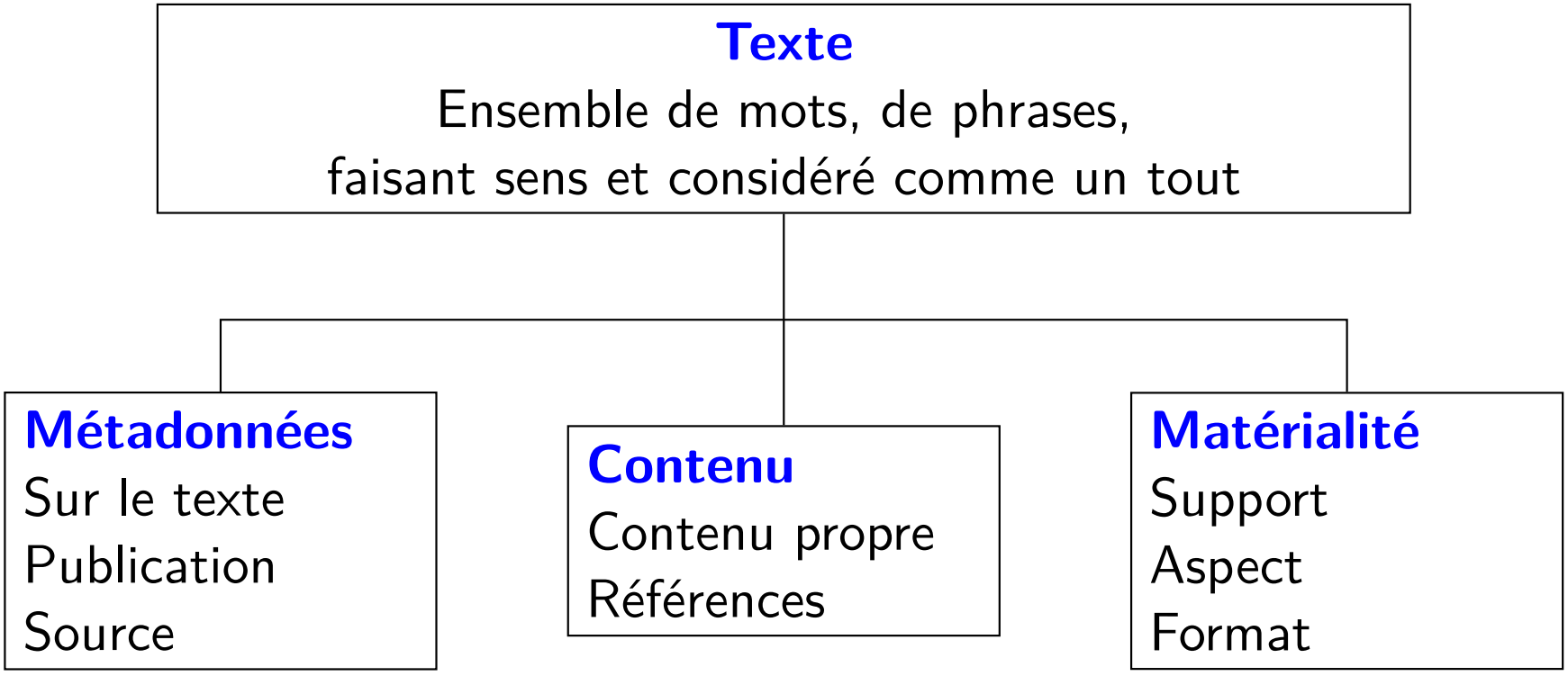
Objectifs du cours

- Comprendre ce qu'est le balisage d'un texte
- Savoir utiliser Markdown et transformer un fichier `.md` au format voulu (Pandoc) (syntaxe de base)
- Comprendre les enjeux liés aux XML
- Savoir lire et composer un document XML de base
- Connaître les technologies liées à XML
- Apprendre les éléments de base dans les langages TEI et EAD
- Première approche des méthodes permettant exploiter les documents XML (XSLT, XPath)

6 séances de deux heures :

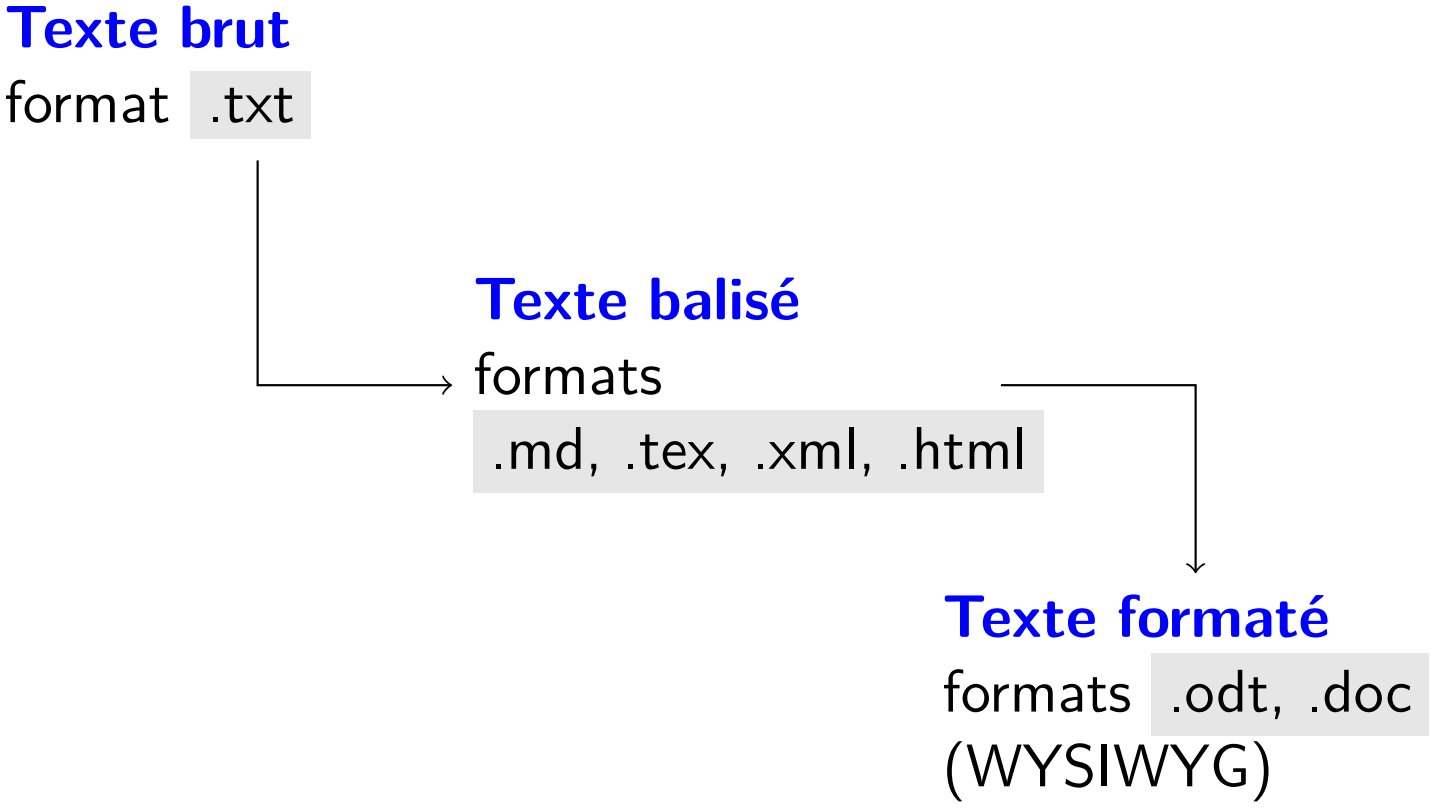
- Une séance (et un peu plus...) sur Markdown et Pandoc
- Deux séances sur le langage et les schémas XML
- Deux séances sur l'édition scientifique avec XML-TEI
- Une séance sur XPath et XSLT

Qu'est-ce qu'un texte ?



Qu'est-ce qu'un texte ?

Le texte informatisé



Qu'est-ce qu'un texte ?

Texte brut

```
Le 15 septembre 1840, vers six heures du matin, la
↪  Ville-de-Montereau, près de partir, fumait à gros
↪  tourbillons devant le quai Saint-Bernard.
```

Texte balisé en XML

```
<paragraphe>Le <date>15 septembre 1840</date>, vers
↪  six heures du matin, la
↪  <nomBateau>Ville-de-Montereau</nomBateau>, près
↪  de partir, fumait à gros tourbillons devant le
↪  quai
↪  <nomLieu>Saint-Bernard</nomLieu>.</paragraphe>
```

Texte formaté

Le 15 septembre 1840, vers six heures du matin, la *Ville-de-Montereau*, près de partir, fumait à gros tourbillons devant le quai Saint-Bernard.

Les langages à balises

- Markdown
- LaTeX
- html
- XML

→ **Mise en forme** (typographique) *vs* **mise en sens** (sémantique)

Les langages à balises

- Markdown
- LaTeX
- html
- XML

Texte balisé en markdown

```
# Première partie
## I
Le 15 septembre 1840, vers six heures du matin, la
 ↪  *Ville-de-Montereau*, près de partir, fumait à
 ↪  gros tourbillons devant le quai Saint-Bernard.
```

Les langages à balises

- Markdown
- LaTeX
- html
- XML

Texte balisé en LaTeX

```
\part{Première partie}
\chapter{I}
Le 15 septembre 1840, vers six heures du matin, la
 ↪  \emph{Ville-de-Montereau}, près de partir, fumait
 ↪  à gros tourbillons devant le quai Saint-Bernard.
```

Les langages à balises

- Markdown
- LaTeX
- html
- XML

Texte balisé en html

```
<h1e>Première partie</h1>
<h2>I</h2>
<p>Le 15 septembre 1840, vers six heures du matin, la
↪  <i>Ville-de-Montereau</i>, près de partir, fumait
↪  à gros tourbillons devant le quai
↪  Saint-Bernard>.</p>
```

Les langages à balises

- Markdown
- LaTeX
- html
- XML

Texte balisé en XML

```
<titrePartie>Première partie</titrePartie>
<numeroChapitre>I</numeroChapitre>
<paragraphe>Le <date>15 septembre 1840</date>, vers
↪  six heures du matin, la
↪  <nomBateau>Ville-de-Montereau</nomBateau>, près
↪  de partir, fumait à gros tourbillons devant le
↪  quai
↪  <nomLieu>Saint-Bernard</nomLieu>.</paragraphe>
```

Les langages à balises

Des langages différents pour des utilisations différentes

Markdown	Un balisage léger pour produire aisément des documents convertibles en différents format
LaTeX	Un balisage plus complexe et très personnalisable pour produire des pdf d'une très grande qualité typographique
XML	Un langage orienté vers le balisage sémantique, pour stocker et interroger des données
html	Un langage pour les pages web

Deux concepts-clefs

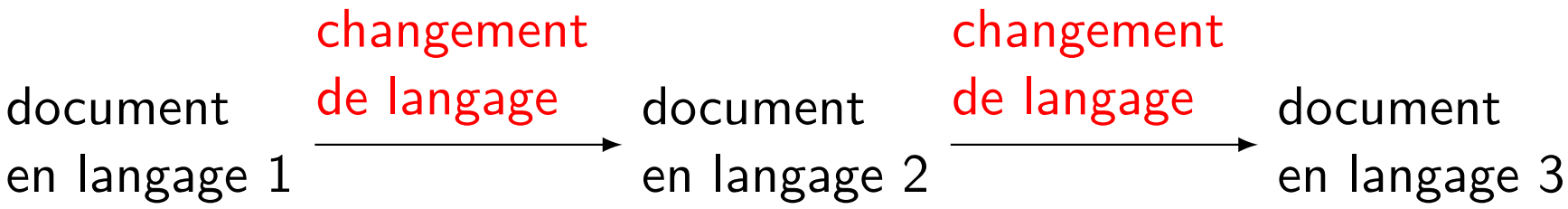
Pérennité

Besoin de formats qui durent dans le temps pour ne pas avoir sans cesse à convertir les fichiers au fil du temps.

Interopérabilité

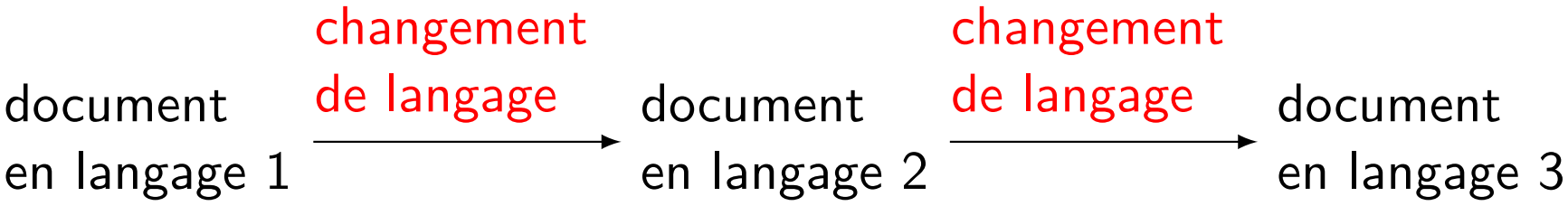
Besoin de formats interopérables, c'est-à-dire qui puissent être échangés et réutilisés, également pour ne pas avoir à convertir sans cesse ses données

Pérennité

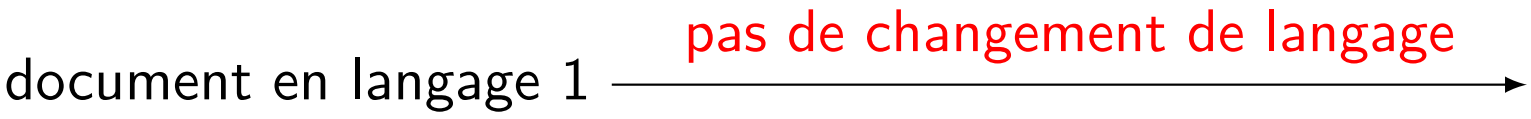


Lorsqu'un langage n'est pas pérenne...

Pérennité

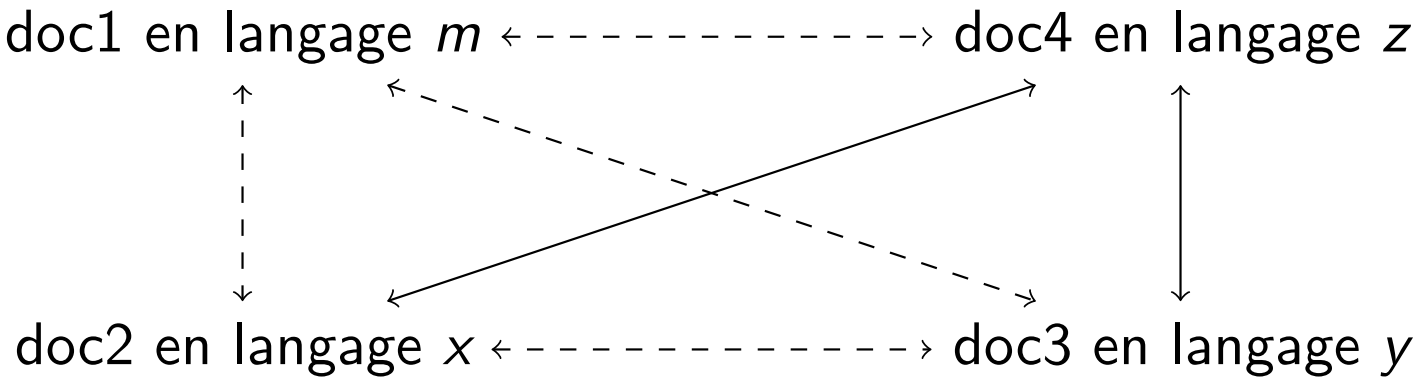


Lorsqu'un langage n'est pas pérenne...



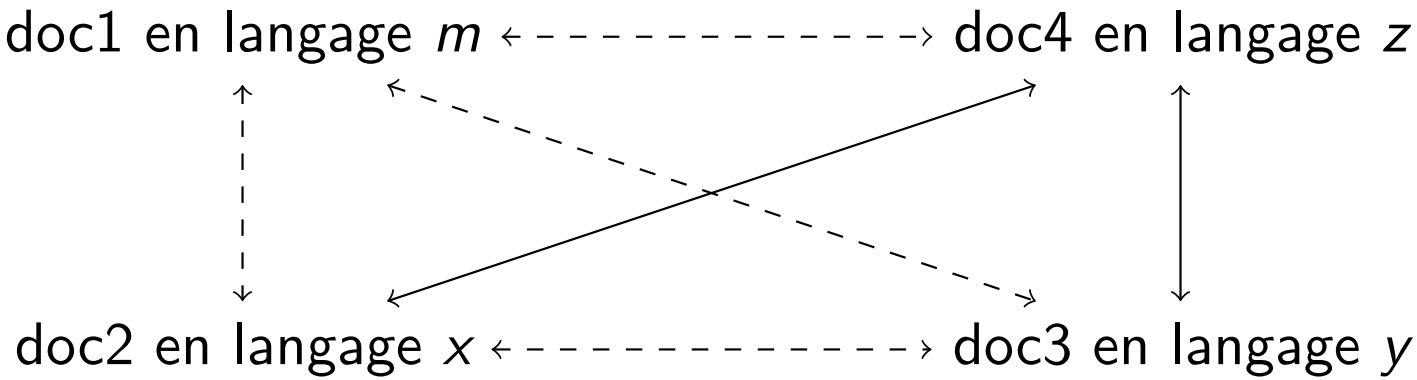
Lorsqu'un langage est pérenne...

Interopérabilité

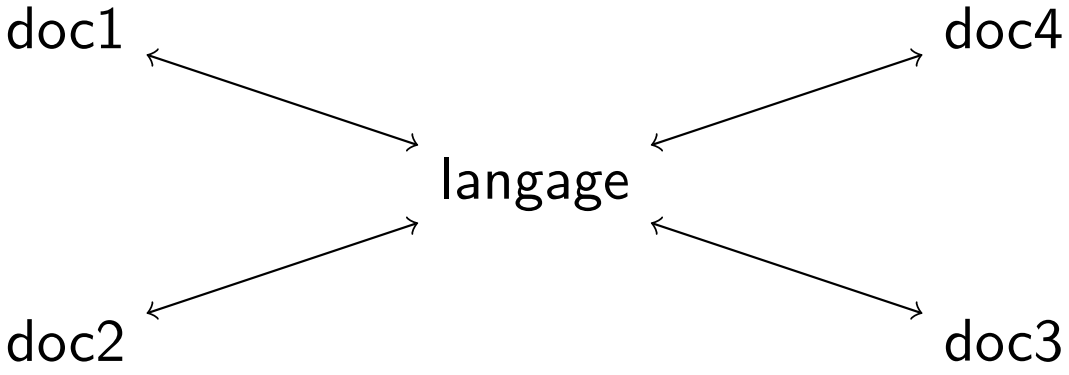


Quand il n'y a pas de langage interopérable...

Interopérabilité



Quand il n'y a pas de langage interopérable...



Quand il y a un langage interopérable

Structurer un texte avec Markdown. Présentation

Présentation de Markdown

- Langage de balisage léger
- Créé par Aaron Swartz et John Gruber en 2004 dans le but de produire facilement du code HTML
- Marquage sémantique du texte explicite et simple
- Fichiers au format `.md` : du texte brut

Avantages de Markdown

- pérennité et, en grande partie, interopérabilité
- facile à lire et à écrire
- séparation fond/forme
- texte brut, donc très léger ; plus écologique.
- *Single Source publishing* : avec un même fichier source, on peut produire des documents dans différents formats (ex : HTML, ePub, pdf, latex, diapo,...).
- De plus en plus de plateformes utilisent markdown. Exemple : GitHub
- **Compatible avec la recherche en SHS** :

Avantages de Markdown

- pérennité et, en grande partie, interopérabilité
- facile à lire et à écrire
- séparation fond/forme
- texte brut, donc très léger ; plus écologique.
- *Single Source publishing* : avec un même fichier source, on peut produire des documents dans différents formats (ex : HTML, ePub, pdf, latex, diapo,...).
- De plus en plus de plateformes utilisent markdown. Exemple : GitHub
- **Compatible avec la recherche en SHS** :
  - format libre : la rédaction de nos articles, ouvrages, etc, ne dépendent pas de formats propriétaires
  - stable, partageable et adapté à nos besoins (facilité pour produire des notes de bas de pages, citer des références bibliographiques, etc)
  - $\rightarrow$ permet la **réappropriation de son outil de travail** par le chercheur

Dans quel contexte utiliser Markdown ?

- prises de note
- rédaction de cours, de wiki
- rédaction d'articles, de livres
- rédaction d'un *readme* dans le cadre d'un dépôt Git
- ...

Les « saveurs » (*flavors*) de Markdown

La simplicité de Markdown en est parfois la limite : beaucoup moins modulable que LaTeX, par exemple

→ Certains programmes ou plateformes ont ajoutés des extensions à Markdown pour l'adapter à des besoins spécifiques

Avantages Plus de souplesse que la syntaxe de base

Inconvénients On s'éloigne de l'interopérabilité et de la légèreté initiale

Les « saveurs » (*flavors*) de Markdown

Exemple de « saveurs », ou spécifications, de Mardown :

- CommonMark
- Pandoc
- GitHub Flavored Markdown
- MultiMarkdown
- ...

Utiliser Markdown : quelques outils

Éditeurs en ligne

- Dillinger (`https://dillinger.io/`)
- StackEdit
- Stylo (un outil d'Huma-Num)
  (`http://stylo-doc.ecrituresnumeriques.ca/fr/`)

Éditeurs en local

- éditeurs spécifiquements dédiés à markdown :
  - ZettIr
  - MarkText
  - Ghost
  - Obsidian, ...
- éditeurs multi-langages :
  - Visual code editor
  - Vim
  - Emacs, ...

Pandoc

Outil de trandsformation du format `.md` vers d'autres formats.

`https://pandoc.org/`

Premiers pas en Markdown

Paragraphes et espaces

Exercice

- Dans l'éditeur en ligne Dillinger, coller le texte du fichier exercice.md (**https ://dillinger.io/**)
- Compléter les métadonnées dans l'en-tête YAML (utilisé par Pandoc, par exemple pour donner un titre au contenu, pour chercher une bibliographie, etc.)
- Ajouter des lignes blanches entre les paragraphes
- Ajouter de simples retours à la ligne au sein des paragraphes
- Ajouter des espaces entre les mots
- Exporter en pdf et observer

Structurer son texte

Résumé :

- Une ou plusieurs lignes blanches $\rightarrow$ paragraphe
- Un ou plusieurs espaces $\rightarrow$ un espace
- Retour chariot : pas de changement de paragraphe
- Les métadonnées ne sont pas forcément visibles dans le nouveau format

Structurer son texte

Ajouter des niveaux de titre

```
# Titre de niveau 1
## Titre de niveau 2
### Titre de niveau trois
... etc
```

Équivalent pour les titres de niveau 1 et 2 :

```
Titre de niveau 1
=====
Titre de niveau 2
-----
```

nb Ne pas oublier l'espace entre le # et le titre

Exercice : Ajouter des titres de niveau 1, 2 et 3

Structurer son texte

Les listes

Listes non numérotées : trois marqueurs possibles, qui peuvent s'utiliser indifféremment - + *

Pour les listes imbriquées, faire précéder de quatre espaces ou une tabulation

```
- un item
+ un autre
* un troisième
    + item d'une liste imbriquée
    - second item
```

- **nb** Ne pas oublier l'espace entre le marqueur et le contenu de l'item.
- Sauter une ligne avant et surtout après vos listes.

Exercice : mettre des items à la liste des courses. Ajouter une liste imbriquée

Structurer son texte

Les listes

Listes numérotées : un nombre suivi d'un point ou d'une parenthèse fermante (ne pas changer). On peut employer n'importe quel nombre.

```
1. un item
1. un autre
3. un troisième
    1) item d'une liste imbriquée
    1) second item
```

nb Ne pas oublier l'espace entre le point ou la parenthèse et le contenu de l'item

Exercice : ajouter une liste imbriquée numérotée

Les listes

Liste à cocher

```
- [ ] item
- [x] item déjà coché
```

Testez !

Les citations

Blocs de citation

```
> Ceci est une citation
>
> Sur plusieurs paragraphes
```

nb Il faut laisser une ligne vide avec chevron pour un saut de paragraphe dans la citation

Exercice : Baliser les citations dans le texte

Les citations

Citer du code

- entre apostrophes
- bloc de code : encadré par trois apostrophes, avec retour à la ligne

```
`...du code`
```

````
```
Un bloc de code plus long
```
````

Exercice : baliser le code dans le texte

Mise en forme

Tester quelques éléments de mise en forme :

Emphases

```
*texte en italique*
**texte en gras**
***texte en gras et italique***
~~barré~~
```

Indices et exposants

```
CO^2^
H~2~O
```

Lignes intercalaires

```
***
ou
---
```

Notes, tableaux, images et liens

Note de bas de page

```
Du texte^[une note "inline"]

Du texte avec une note par référence[^1] et une
↪  seconde[^3]

[^1]: L'importance est d'être cohérent entre l'appel
↪  de note et la note
[^3]: La numérotation se fait automatiquement
```

nb Pour les notes par référence : ne pas oublier de mettre un espace après les deux points, ni de séparer le corps du texte et le contenu de la note par une ligne blanche

Tester !

Notes, tableaux, images et liens

Tableaux

```
|**gauche**|**centré**|**droite**|
|:---|:---:|---:|
|cellule 1.1|cellule 2.1|cellule 3.1|
|cellule 1.2|cellule 2.2|cellule 3.2|
|cellule 1.3|cellule 2.3|cellule 3.3|
: Titre du tableau
```

nb : En modifiant le nombre de tirets, on peut modifier la largeur des colonnes, lorsqu'on exporte avec Pandoc

Exercice : produire le tableau suivant :

Fruits	**Légumes**
Poires	Courgettes
Pommes	Tomates

nb : Ajoutez un titre :il n'est pas exporté. Nous verrons comment mieux gérer l'exportation vers d'autres formats avec Pandoc.

Notes, tableaux, images et liens

Liens et images

```
[Texte du lien](http://lien.fr)
![légende de l'image](chemin vers l'image: fichier ou url)
```

Exercice : Créer un lien vers le site de votre choix et faire apparaître une image

Les limites de l'éditeur en ligne

- on ne peut pas importer d'images sans passer par dropbox, github, etc
- on ne peut pas importer de bibliographie
- on ne maîtrise pas l'outil de transformation (pandoc)

Convertir le format Markdown : Pandoc

Qu'est-ce que Pandoc

- Un programme de conversion entre formats de balisage.
- Reconnaît et sait transformer des formats comme `.html`, `.xml`, `.tex`, `.docx`, `.odt`, `.pptx`, `.bib`, `.csl`, `.json` ...

Utilisation de base de pandoc

- Créer un dossier dédié au cours, y importer le fichier markdown créé avec Dillinger sous le nom `exercice.md`
- Ouvrir un terminal et se placer dans le répertoire (dossier) que l'on vient de créer
- Taper : `pandoc exercice.md -o exercice.html`
- Ouvrir le fichier obtenu avec un navigateur
- Refaire la même opération avec les changements nécessaires pour obtenir un fichier pdf
- Observez $\rightarrow$ Les métadonnées sont utilisées dans le pdf ainsi créé

Pour la semaine prochaine

- Testez et choisissez un éditeur markdown adapté à votre système d'exploitation.
- Entraînez-vous à taper en markdown en utilisant au moins les niveaux de titre. Exemple : prise de notes pendant un cours, notes de lecture sur un article,...
- Dans Zotéro, si vous n'en avez pas, importez des références bibliographiques.

Quelques références sur Markdown et Pandoc

- https://e-publish.uliege.be/md/
- https://programminghistorian.org/fr/lecons/redaction-durable-avec-pandoc-et-markdown
- https://www.jdbonjour.ch/cours/markdown-pandoc/
- https://eveille.hypotheses.org/975
- https://he-arc.github.io/rapport-technique/rapport.pdf