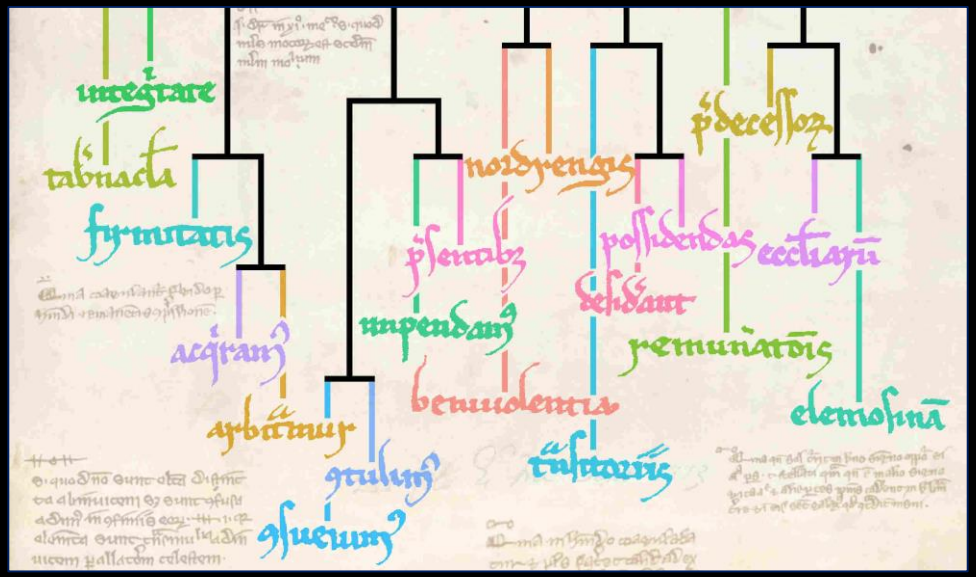
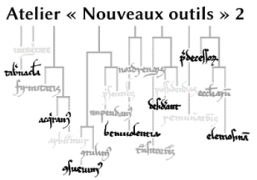
Du parchemin à la fouille de données

Nouveaux outils pour la création, la formalisation et l'analyse des corpus médiévaux

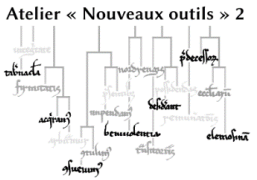


Paul Bertrand
Étienne Cuvelier
Sébastien de Valeriola
Nicolas Perreaux
Nicolas Ruffini-Ronzani



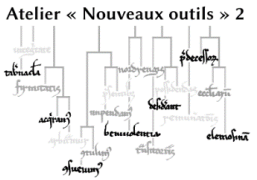
1. Fonctions R à utiliser

Tâche à effectuer	Fonction (du paquet `stringr`)
Détecter un motif	`str_detect()` `str_which()`
Extraire un motif	`str_extract(), str_extract_all()` `str_subset()`
Localiser un motif dans une chaîne	`str_locate()` `str_locate_all()`
Remplacer un motif par un autre	`str_replace()` `str_replace_all()`
Diviser une chaîne à l'emplacement d'un motif	`str_split()`
Supprimer les espaces blancs au début et à la fin d'une chaîne	`str_trim()`



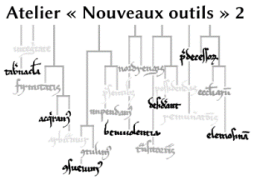
2. Séquences d'échappement

\n	\r	\t	\'	\"
nouvelle ligne	retour de chariot	tabulation	apostrophe	guillemet



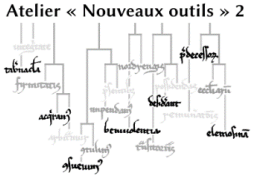
3. Opérateurs

.	[...]	[^...]	$[x_1-x_2]$	|
n'importe quel caractère	un des caractères écrits entre les crochets	un caractères qui n'est pas écrit entre les crochets	intervalle des caractères entre x_1 et x_2	OU logique



4. Quantificateurs

*	+	?	{n}	{n,}	{n,m}
au moins 0 fois	au moins 1 fois	au plus 1 fois	exactement n fois	au moins n fois	entre n et m fois

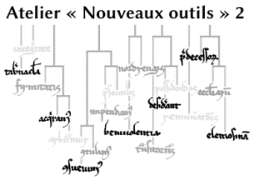


5. Comportement des quantificateurs

comportement	explication	comment l'obtenir ?
gourmandise (*greediness*)	répète le motif autant que possible	comportement par défaut
paresse (*laziness*)	répète le motif aussi peu que possible	ajouter ? après le quantificateur

6. Classes de caractères

[:digit:] ou \d	\D	[:lower:]	[:upper:]	[:alpha:]	[:alnum:]
chiffres (≡ [0-9])	non-chiffres (≡ [^0-9])	lettres minuscules (≡ [a-z])	lettres majuscules (≡ [A-Z])	caractères alphabétiques (≡ [A-z])	caractères alphanumeriques (≡ [A-z0-9])
\w	\W	[:xdigit:]	[:blank:]	[:space:]	\s
caractères "mots" (≡ [A-z0-9_])	caractères "non-mots" (≡ [^A-z0-9_])	chiffres hexadécimaux (≡ [0-9A-Fa-f])	caractères blancs (espaces et tabulations)	caractères d'espacement (tabulation, nouvelle ligne, etc.)	espace
\S	[:punct:]	[:graph:]	[:print:]	[:cntrl:]	
non-espace	ponctuation (! " # $ % & ’ ( ) * + , - . / : ; < = > ? @ [ ] ^ _ ` { | } ~)	caractères graphiques (≡[[:alnum:] [:punct:]])	caractères visibles (≡ [[:alnum:] [:punct:]] \\s)	caractères de contrôle (\n, \r, etc.)	



7. Position du motif dans la chaîne

^	$	\b	\B
début de la chaîne	fin de la chaîne	caractère vide désignant le début d'un mot	caractère vide désignant la fin d'un mot

8. Retro-références (*backreferences*)

`(...)`	`(?:...)`
groupe capturant	groupe non-capturant