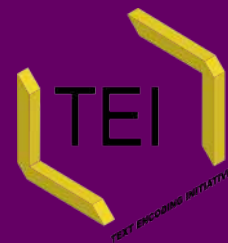


# Введение в TEI



А.М. Лаврентьев

Лаборатория ICAR – Национальный центр  
научных исследований Франции и Лионский  
университет

El'Manuscript 2012

Петрозаводск, 4.09.2012

# План

- Стандарт Unicode (?)
- Стандарт XML (?)
- Теория TEI
  - Краткая история (?)
  - Документация: TEI Guidelines
  - Персонализация: ODD
- Практика TEI
  - Критические издания и корпуса текстов
  - Стратегия разметки документа в TEI
  - Совместимость с лингвистической разметкой
  - Примеры спецификации

# Стандарт Юникод

- <http://unicode.org>
- Существует с 1991 г.
  - актуальная версия 6.1 (2012)
- Позволяет представить *почти* все знаки письменных языков
  - 1 112 064 кодовых позиций
- Решает проблемы совместимости различных систем письма
- Поддерживается большинством современных операционных систем и программных продуктов

# Стандарт Юникод

## ■ Состоит из

- универсального набора символов UCS
- семейства кодировок UTF
  - UTF-8 - наиболее распространенная
- обозначение кода
  - U+xxxx, U+xxxxx или U+xxxxxx
- 17 «плоскостей» по  $2^{16}$  (65536) символов
  - xxxx - базовая, включает зону частного использования
  - 1-я плоскость - для исторических систем письма
  - 15-я и 16-я - для частного использования

# Стандарт Юникод

- Группы символов
  - буквы
  - цифры
  - знаки пунктуации
  - специальные знаки

# XML и TEI

- Лекции Лу Бернарда (Казань, 2008)
  - Краткое введение в XML
  - Краткое введение в TEI
  - Основные структурные элементы TEI
    - <http://tei.oucs.ox.ac.uk/Oxford/2008-08-kazan/>

# Проекты, использующие TEI

The screenshot shows a Mozilla Firefox browser window with the address bar displaying <http://www.tei-c.org/Activities/Projects/>. The page title is "TEI: Projects Using the TEI". The website has a blue header with the TEI logo and the text "< Text Encoding Initiative >". Below the header is a navigation menu with links: Home, Guidelines, Activities, Tools, Membership, Support, About, News, and Online Store. A breadcrumb trail shows "Home -> Activities -> Projects". On the left side, there is a sidebar with links: TEI Council, Workgroups, Special Interest Groups (SIGs), Projects, and jTEI - Journal of the TEI. The main content area is titled "Projects Using the TEI" and contains a paragraph explaining that the following is a list of projects that use the TEI encoding scheme. It invites users to add their project to the list by filling out a [new project form](#) and to keep the pages up to date by reviewing their project description and sending updates to [web@tei-c.org](mailto:web@tei-c.org). Below this is a section titled "List of Projects" which contains a bulleted list of project links:

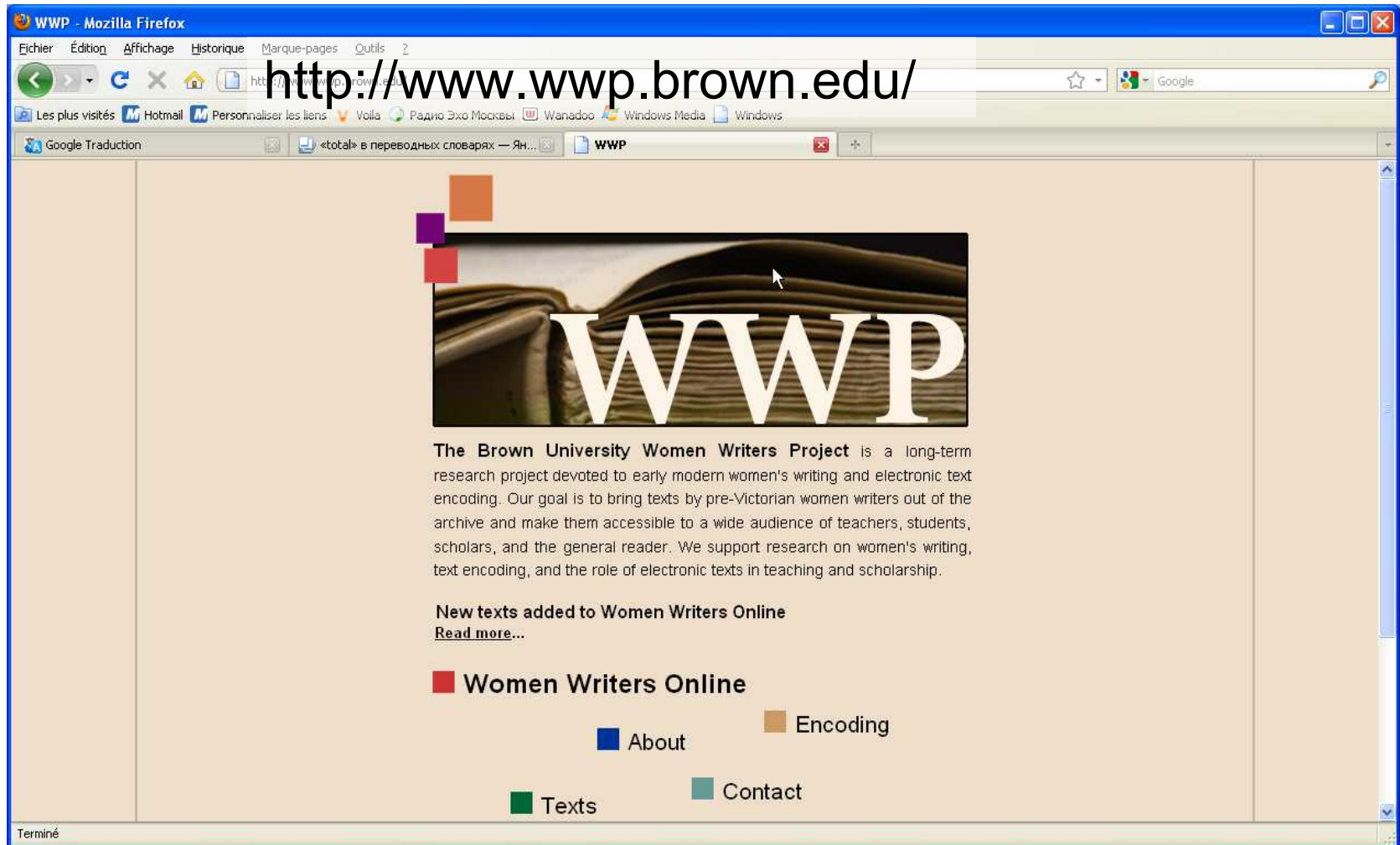
- [African American Women Writers of the 19th Century](#)
- [African Languages Lexicon Project \(ALLEX\)](#)
- [Alex Catalogue of Electronic Texts](#)
- [American Memory from the Library of Congress](#)
- [American Verse Project](#)
- [The Anglo-Saxon Poetry Project](#)
- [Aphrodisias in Late Antiquity \(2004\)](#)
- [Archimedes Palimpsest Project](#)
- [ATLAS : ATLA Series](#)
- [Autour des Res Gestae Divi Augusti](#)
- [Berardier.org: Édition électronique de Bérardier de Bataut, Essai sur le récit \(1776\)](#)
- [Boccaccio's Decameron](#)

The browser's status bar at the bottom shows the URL <http://www.tei-c.org/Activities/Projects/an01.xml>.

# Early Americas Digital Archive



# Women Writers Project



# «Артамен, или Великий Кир »

Artamène ou le Grand Cyrus - Mozilla Firefox

http://www.artamene.org/

Les plus visités M Hotmail M Personnaliser les liens V Voila Радно Эхо Москвы W Wanadoo Windows Media Windows

Google Traduction «total» в переводных словарях — Ян... Early Americas Digital Archive Artamène ou le Grand Cyrus

ARTAMÈNE OU LE GRAND CYRUS

PROJETS  
CPDM  
LE RÉGNE D'ASTRÉE  
MOLIÈRE 21

NAVIGATION  
• RECHERCHE DE MOTS  
• RECHERCHE DE PAGES  
• TÉLÉCHARGEMENT

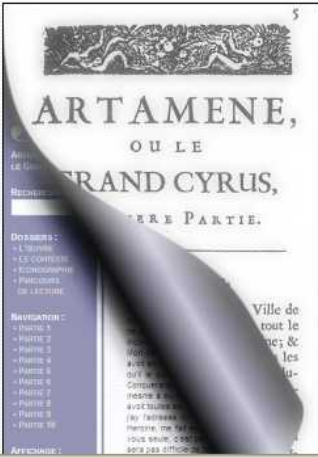
TEXTE  
• SYNOPSIS  
• PARTIE 1  
• PARTIE 2  
• PARTIE 3  
• PARTIE 4  
• PARTIE 5  
• PARTIE 6  
• PARTIE 7  
• PARTIE 8  
• PARTIE 9  
• PARTIE 10  
• ILLUSTRATIONS

ENCYCLOPÉDIE  
• SOMMAIRE  
• NOUVEAUTÉS

DOCUMENTS

Présentation

Le site "Artamène" offre l'intégralité du texte du plus long roman de la littérature française, *Artamène ou le Grand Cyrus* (1649-1653) de Madeleine de Scudéry (13 095 pages dans son édition originale, 7443 dans cette édition en ligne).



Madeleine et Georges de Scudéry  
*Artamène ou le Grand Cyrus*

accueil | projet | œuvre | édition | contacts

Lire le *Grand Cyrus*

*Artamène ou le Grand Cyrus* est-il véritablement cette œuvre d'une lecture ardue, voire impossible, que les commentateurs se sont acharnés depuis trois siècles à dénigrer, même lorsqu'ils lui reconnaissent des qualités ?

Le plus long roman français pêche-t-il vraiment par démesure, par profusion, par redondance ?

Comment dès lors expliquer le succès immense de l'ouvrage auprès du public de l'époque ?

Et si tout cela n'était qu'un problème de " lecture " ?

La lecture intime et linéaire, telle que nous l'appliquons à tous les textes littéraires, est-elle vraiment le moyen approprié d'aborder le *Grand Cyrus* ? Œuvre élaborée au sein d'un salon, le roman n'est-il pas lui aussi conçu en fonction du mode de consommation des genres mondains, la lecture à haute voix dans un contexte social interactif : une lecture par extraits renvoyant à d'autres extraits, devant un auditoire choisi, dont les membres sont capables de mettre en relation les composantes de l'ouvrage entre elles, de les commenter et de suggérer diverses pistes dans le développement de l'intrigue ?

Redonner une chance au *Grand Cyrus*, c'est peut-être retrouver ce mode de consommation originel.

A défaut de reconstituer le modèle social du salon, ne peut-on au moins en retrouver l'esprit dans les possibilités d'accès au texte et les facultés interactives propres à Internet ?

Hypertiens, c'est-à-dire renvois...

Parcours de lecture non linéaire au gré des suggestions de l'auditoire, en d'autres termes...

http://www.artamene.org/synopsis.php

# Параллельное издание буддистского текста Samyukta Agama

http://buddhistinformatics.chibs.edu.tw/BZA/

漢文古籍譯註與數位編輯的研究——  
以巴利語與漢文《別譯雜阿含經》(T.100)的版本比對與英譯為例  
A Digital Comparative Edition and Partial Translation of the Shorter  
Chinese Samyukta Āgama (T.100)

The Digital Comparative Edition of the Bieyi za ahan jing 別譯雜阿含經 (BZA) is a project undertaken by the Dharma Drum Buddhist College 法鼓佛教研修學院 and funded by the Chiang Ching-kuo Foundation for International Scholarly Exchange 蔣經國國際學術交流基金會.

This comparative digital edition:

- provides new punctuation for the BZA and the Za ahan jing 雜阿含經 (ZA) sutras
- corrects and documents mistakes in previous editions
- distinguishes and visualizes parallel and non-parallel passages between the BZA and other Chinese and Pāli versions, enabling the user to conveniently compare the different texts of a cluster
- refines and expands the contents of the 364 text clusters
- provides an annotated English translation of selected sections of the BZA
- enables statistical linguistic analysis by creating aligned parallel corpora (not online)
- is extensible and allows for further material to be added

Choose a Cluster

Search

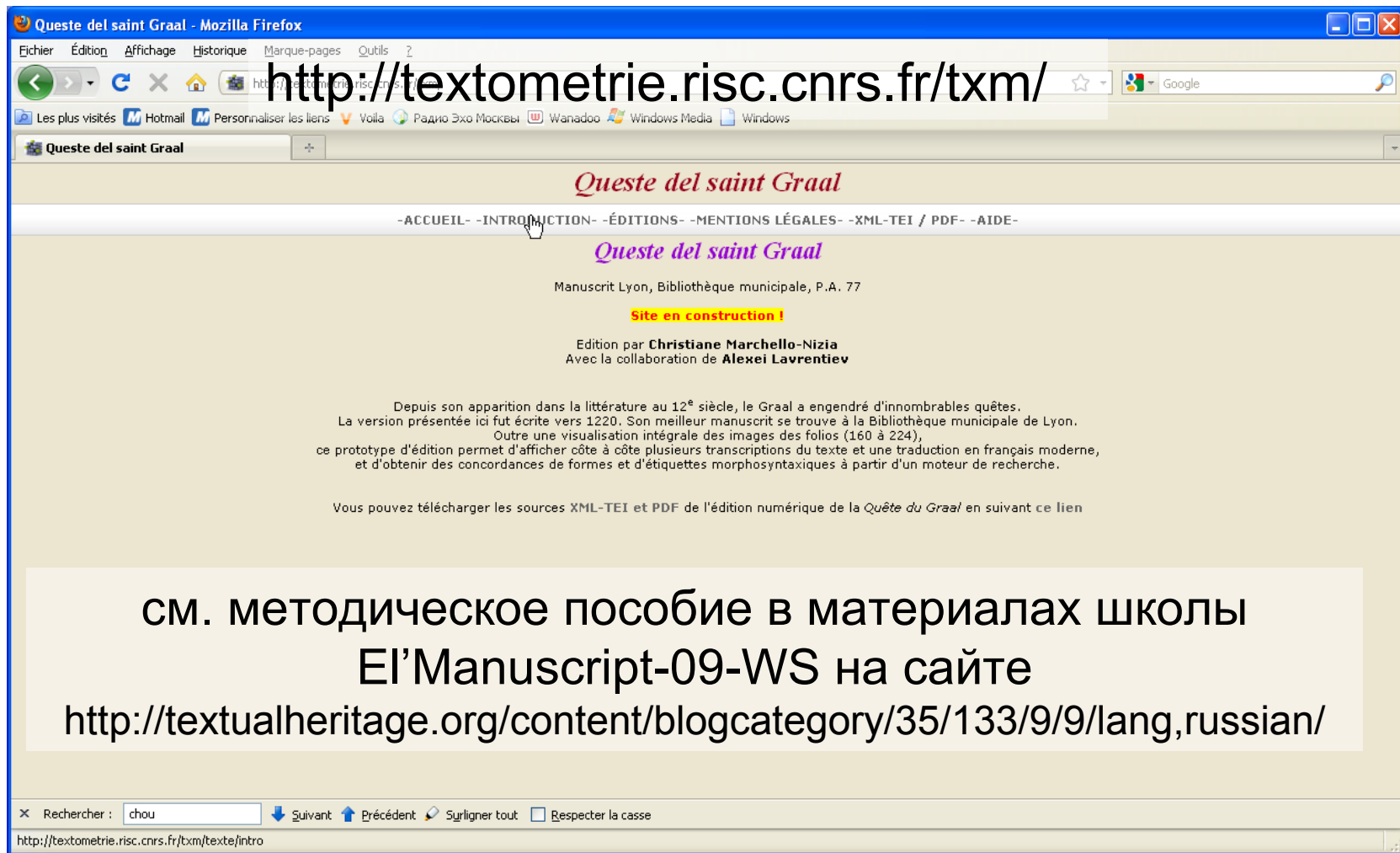
BZA Sutras in Original Order

BZA Topics

Abbreviations

http://buddhistinformatics.ddbc.edu.tw/solr/BZA/

# Проект *Graal*



# Проект Graal

Queste del saint Graal: Introduction - Mozilla Firefox

Echier Édition Affichage Historique Marque-pages Outils ?

http://textometrie.risc.cnrs.fr/txm/texte/intro

Les plus visités Hotmail Personnaliser les liens Voila Радио Эхо Москвы Wanadoo Windows Media Windows

Queste del saint Graal: Introduction

## Queste del saint Graal

-ACCUEIL- -INTRODUCTION- -ÉDITIONS- -MENTIONS LÉGALES- -XML-TEI / PDF- -AIDE-

### Sommaire

- Présentation
- 1. Une édition numérique
- 2. Balisage XML-TEI
- 3. Le 'Lancelot-Graal'
- 4. Résumé de la 'Queste'
- 5. Origines du 'Graal'
- 6. Manuscrits
- 7. Éditions antérieures
- 8. Choix du manuscrit
- 9. Principe de la transcription
- 10. Voyelles et consonnes
- 11. Signes diacritiques
- 12. Abréviations
- 13. Segmentation des mots
- 14. Corrections éditoriales
- 15. Corrections scribales
- 16. Majuscules
- 17. Ponctuation
- 18. Structuration du texte
- 19. Langage du copiste
- 20. Étiquettes morphosyntaxiques
- 21. Bibliographie
- 22. Études annexes

### Brève présentation de cette édition de la *Queste del saint Graal*, roman en prose du 13 e siècle

L'édition de la *Queste del saint Graal* présentée ici est **un objet purement numérique** (cf. section 1), caractérisé par sa nature multi-facettes : il ne pourrait en exister UNE version imprimée unique, mais chaque utilisateur peut imprimer ou télécharger la version qui correspond à son besoin ponctuel. C'est un objet dynamique par les possibilités qu'il offre d'un affichage en deux colonnes au choix, de liens multiples, et d'étiquettes ou de notes affichées selon les besoins ; en constant enrichissement, et interactif, son utilisation optimale est numérique. Et **le formatage XML avec des balises qui suivent les recommandations de la TEI** vous permettent, dans les conditions juridiques de « *Creative Commons* », de charger tout ou partie des éléments ici présentés afin de les intégrer à votre propre environnement ou à vos propres usages (cf. section 2).

La *Quête du saint Graal* est un roman en prose écrit en France vers 1225-1230, et dont l'auteur est inconnu. Il appartient à **un vaste ensemble de récits** consacrés au roi Arthur, à l'enchantement Merlin, aux chevaliers de la Table Ronde, à l'amour adultère de Lancelot avec la reine Guenièvre, femme du roi Arthur, et à cette aventure chevaleresque et mystique qu'est la « quête » du Graal, qui précède et annonce la fin du monde arthurien, que conte le dernier roman de la série, *La Mort le roi Artu* (cf. section 3). De tous ces récits, la *Quête du saint Graal* est sans doute le plus énigmatique, le plus beau aussi, celui dont on n'épuisera jamais le **sens** (cf. section 4). On en rappellera les **origines** (cf. section 5).

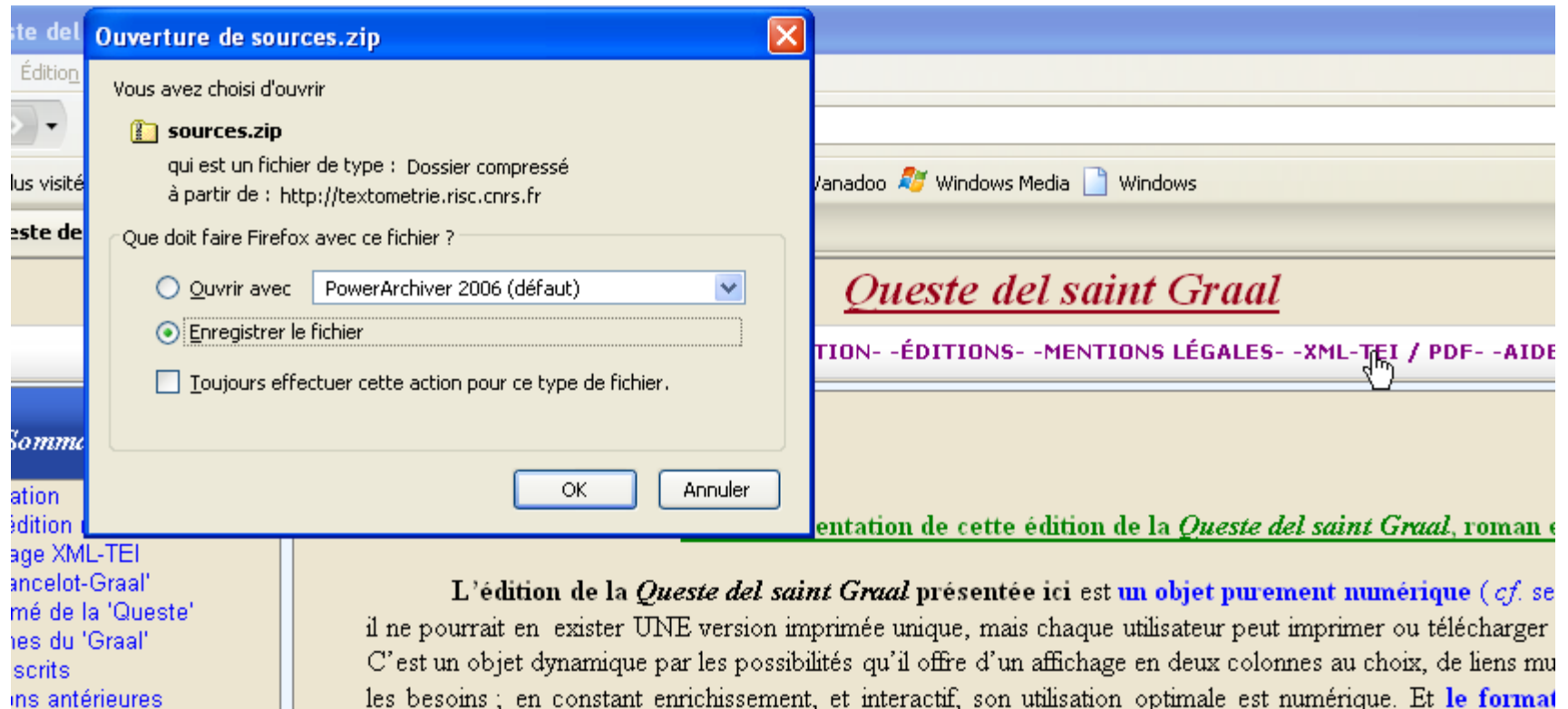
Ce roman célèbre a été transmis par **une cinquantaine de manuscrits** (cf. section 6), et a donné lieu à plusieurs éditions déjà, que nous listons (cf. section 7). Nous éditons ici l'une des versions les plus anciennes, conservée dans un manuscrit qui se trouve à Lyon (manuscrit K, Bibliothèque Municipale, Palais des Arts 77) (cf. section 8). Nous en donnons une édition très fidèle, mais qui, contrairement aux éditions imprimées, peut être visionnée sous trois formes différentes, suivant le niveau de lecture souhaitée : texte 'normal', texte dans lequel les abréviations sont signalées, texte 'facsimilaire' qui tente de restituer toutes les particularités formelles du manuscrit (cf. section 1). Nous avons explicité les principes qui ont guidé notre édition et décrit avec précision **la façon dont nous avons transcrit le manuscrit** (cf. sections 9, 10, 11, 12 et 13), et dont nous en avons corrigé **certaines erreurs** (cf. sections 14 et 15). Nous avons précisé la **structuration** apportée au texte aux divers niveaux, par la ponctuation ou par les lettres rubriquées (cf. sections 16, 17 et 18). Ces

Rechercher : chou

Suivant Précédent Surligner tout Respecter la casse

Terminé

# Проект *Graal*



# Проект *Graal*

The screenshot shows a web browser window titled "Queste del saint Graal - Mozilla Firefox". The address bar displays the URL <http://textometrie.risc.cnrs.fr/txm/texte/quete>. The website's navigation bar includes links: -ACCUEIL- -INTRODUCTION- ÉDITIONS- MENTIONS LÉGALES- XML-TEI / PDF- AIDE-. Below this, there are tabs for MS:COLONNE, FACSIMILÉ (selected), DIPLOMATIQUE, COURANTE, TRADUCTION, and MS:FOLIO. The main content area is split into two columns. The left column shows a facsimile of a manuscript page with a large decorated initial 'D' in blue and red. The right column shows the corresponding modern French translation, with line numbers 5, 10, 15, and 20. The translation text is as follows:

§ 5] Quant li rois fu [\[revenuz\]](#) [d]el mostier, et il vit que Lancelot fu venuz et il ot amené Boort et Lion si lor fet mout grant joie, et dist que bien soient il venuz, et la feste comence par laiencz grant et merueilleuse. car mout sont liez li compaignon de la table reonde de la venue as .ii. freres, et mes sires Gauvains lor demande coment il l'ont puis fet que il se partirent de cort, et il dient : « Bien Dieu merci. » Car il ont toz jorz esté sainz et haitez. - « Certes, fet mes sires Gauvains, ce me plect mout. »

§ 6] Granz est la joie que cil de la cort font a Boort et a Lion, car pieça mes qu'il nes avoient veuz, et li rois comande que les tables soient mises, car il est (p. 5) tens de mengier ce li est avis. « Sire, fait Keus li seneschaux, se vos asseez ja au disnier il m'est avis que vos enfreindroiz la costume de ceainz, car nos avons veu toz jorz que vos a haute feste n'asseiez a table devant que aucune aventure fust en vostre cort avenue voiant toz les barons de vostre ostel. - Certes, fet li rois, Keus, vos dites

At the bottom of the browser window, there is a search bar with the text "Rechercher : chou", navigation buttons for "Suivant" and "Précédent", and a "Vue Simple»" button.

# Проект *Graal*

Queste del saint Graal - Mozilla Firefox

http://textometrie.risc.cnrs.fr/txm/texte/quete

Queste del saint Graal

*Queste del saint Graal*

-ACCUEIL- -INTRODUCTION- -ÉDITIONS- -MENTIONS LÉGALES- -XML-TEI / PDF- -AIDE-

MS:COLONNE FACSIMILÉ DIPLOMATIQUE COURANTE TRADUCTION MS:FOLIO

FACSIMILÉ DIPLOMATIQUE COURANTE TRADUCTION

Quant li rois fu [revenu] [d]el mostier, et il vit  
que Lancelot fu venuz et il ot amené Boort et Lion  
si lor fet mout grant joie, et dist que bien soient  
il venuz, et la feste comence par laienz grant et merveilleuse,  
5 car mout sont liez li compaignon de la  
table reonde de la venue as .ii. freres, et mes sires  
Gauvains lor demande coment il l'ont puis fet que il  
se partirent de cort, et il dient : « Bien Dieu merci. » Car  
il ont toz jorz esté sainz et haitez. - « Certes, fet mes  
10 sires Gauvains, ce me plest mout. »

Grand est la joie que cil de la cort font a Boort

Outils ▾ |< < 160d > >| Vue Simple»

Fonction

Concord

Chercher dans Courante

Chercher Lancelot

Propriété de tri word

Taille contexte gauche 10

Reference : col ☒ line ☒ p:n ☐ s:n ☐

Taille contexte droit 10

Propriétés d'affichage word

pos

Nombre de lignes par page 5

Rechercher : chou

Suivant Précédent Surigner tout Respecter la casse

Terminé

# Проект *Graal*

Queste del saint Graal - Mozilla Firefox

Eichier Édition Affichage Historique Marque-pages Outils ?

http://textometrie.risc.cnrs.fr/txm/texte/quete

Les plus visités M Hotmail M Personnaliser les liens V Voila R Радио Эхо Москвы W Wanadoo Windows Media Windows

Queste del saint Graal

*Queste del saint Graal*

-ACCUEIL- -INTRODUCTION- -ÉDITIONS- -MENTIONS LÉGALES- -XML-TEI / PDF- -AIDE-

MS:COLONNE FACSIMILÉ DIPLOMATIQUE COURANTE TRADUCTION MS:FOLIO

FACSIMILÉ DIPLOMATIQUE COURANTE TRADUCTION

[§ 6] uoir . ceste costume ai ie toz iorz tenue elaten/  
drai tant comie porrai . mesieauoie si grant  
ioie de **lancelot** . edelef coufins qui estoient uenu  
acort sain e haitie qui ne me fouenoit dela  
5 costume . ouoienfouuegne fet .K. .

[§ 7] **E**Ndementres qui parloient ainsi fi ent  
laienz un' ualez qui dist au roi . Sire  
noueles uof apozt mout merueilleuf . Quelef  
fet li rois . Dites les moi tost . Sire la aual defoz  
10 uofre pales a . i perron grant que ie ai ueu  
floter par desus l'eue . Venez le ueoir . car ie fai

[§ 6] uoir, ceste costume ai ie toz iorz tenue **et** la ten-  
drai tant com ie porrai, mes ie auoie si grant  
ioie de **lancelot** **et** de ses cousins qui estoient uenu  
a cort sain **et** haitie qui ne me souenoit de la  
5 costume, or uos en souueigne fet **Keus**.

[§ 7] **E**Ndementres qu'il parloient ainsi si entra  
laienz uns ualez qui dist au roi. Sire  
noueles uos apozt mout merueilleuses. Queles  
fet li rois. Dites les moi tost. Sire la aual desoz  
10 uostre pales a . i perron grant que ie ai ueu  
floter par desus l'eue. Venez le ueoir, car ie sai

Outils ▾

161a

Vue Simple»

Fonction Résult. 0

Référence	Contexte Gauche	Pivot	Contexte Droit
col:161a,line:3	je porrai, mes je avoie si grant joie de	Lancelot	et de ses cousins qui estoient venu a cort sain
col:161a,line:23	Et quant li rois voit ces letres si dist a	Lancelot	: « Biau sire ceste espee est vostre par bon
col:161a,line:38	dou saint Graal . » Quant li rois ot que	Lancelot	n en fera plus si dist a mon seignor Gauvain
col:161b,line:1	saue vostre grace non feré, puis que mes sires	Lancelot	n i velt essayer g i metroie la main por
col:161b,line:8	fet mon comandement . - Mes sires Gauvains, fet	Lancelot	, or sachiez que ceste espee vos touchera encore si

15-19 / 225

Rechercher : chou

Suivant Précédent Surigner tout Respecter la casse

Terminé

# TEI и корпуса текстов

- В корпусной лингвистике использование TEI ограничено
  - недостаточно инструментов, поддерживающих TEI
  - стандарт TEI слишком «гибкий»
  - не хватает стандартных механизмов аннотации
  - трудно совместить глубокую филологическую разметку с лингвистической аннотацией

# TEI и корпуса текстов

## ■ Тем не менее:

### □ используют TEI как «родной формат»

- Британский национальный корпус  
<http://www.natcorp.ox.ac.uk/>
- Польский национальный корпус <http://nkjp.pl/>
- Болгарский национальный корпус  
[http://ibl.bas.bg/BGNC\\_bg.htm](http://ibl.bas.bg/BGNC_bg.htm)
- База средневекового французского  
[http://txm.bfm-corpus.org/bfm/...](http://txm.bfm-corpus.org/bfm/)

### □ предлагают экспорт в формате TEI

- Frantext [http://www.frantext.fr/...](http://www.frantext.fr/)

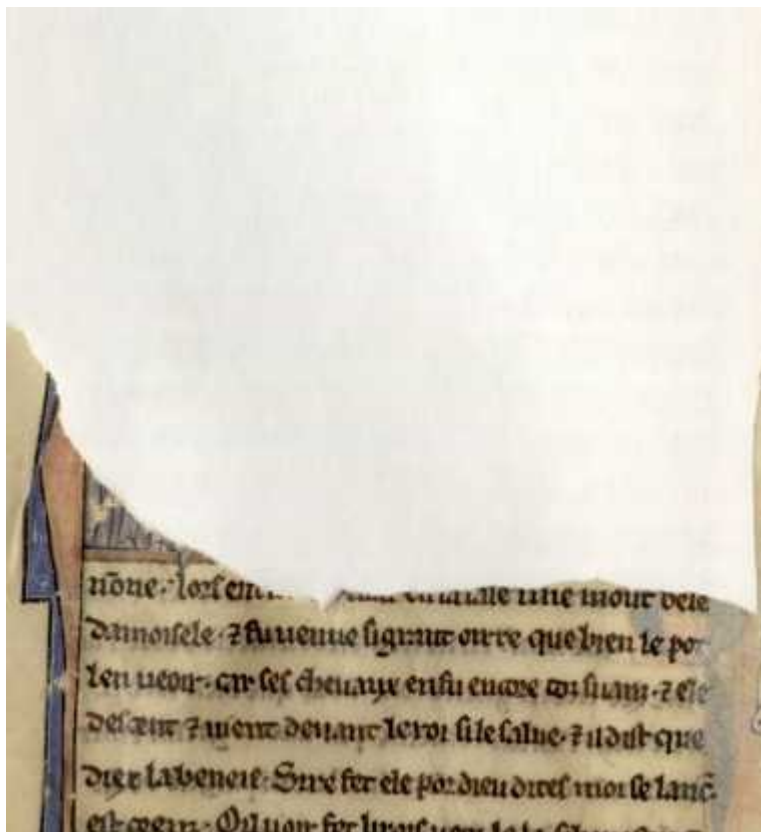
# Стратегия разметки TEI

- Определить цели проекта
  - для кого?
  - основные требования и «бонусы»
- Изучить опыт похожих проектов
  - не нужно изобретать велосипед!
- Определить базовую структуру
  - физическая или логическая?
    - XML не допускает перекрещивания элементов!
- Виды и последовательность разметки
  - правка, варианты, генетическая, лингвистическая

# Стратегия разметки TEI

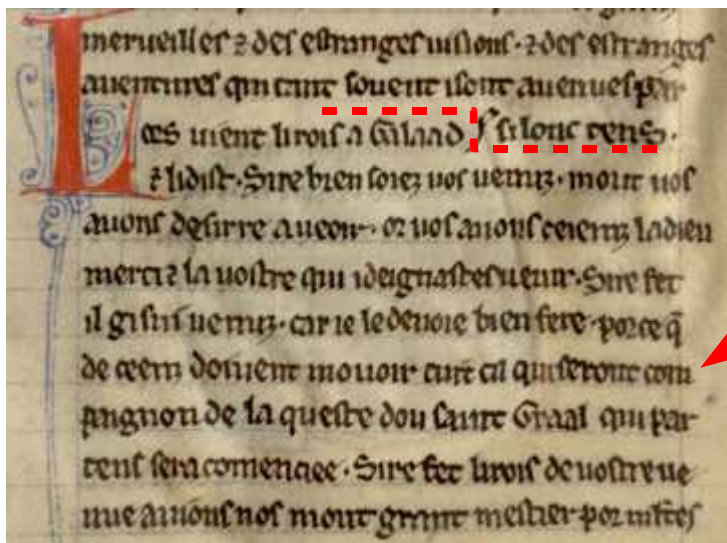
- С помощью ODD
  - Выбрать необходимые модули
  - Удалить ненужные элементы
  - Обосновать принятые решения
- При необходимости обратиться за помощью
  - [TEI-L@LISTSERV.BROWN.EDU](mailto:TEI-L@LISTSERV.BROWN.EDU)
- Собрать метаданные
  - `teiHeader`
- Подготовить стилевую таблицу CSS для редактирования

# Примеры перекрещивания структур



- Лингвистическая / редакторская
  - `<supplied>` A la veille de la Pentecoste [...] tables a heure de`</supplied>` nonne lors en`<supplied>`tra a cheval`</supplied>``<unclear>`en la s`</unclear>`ale

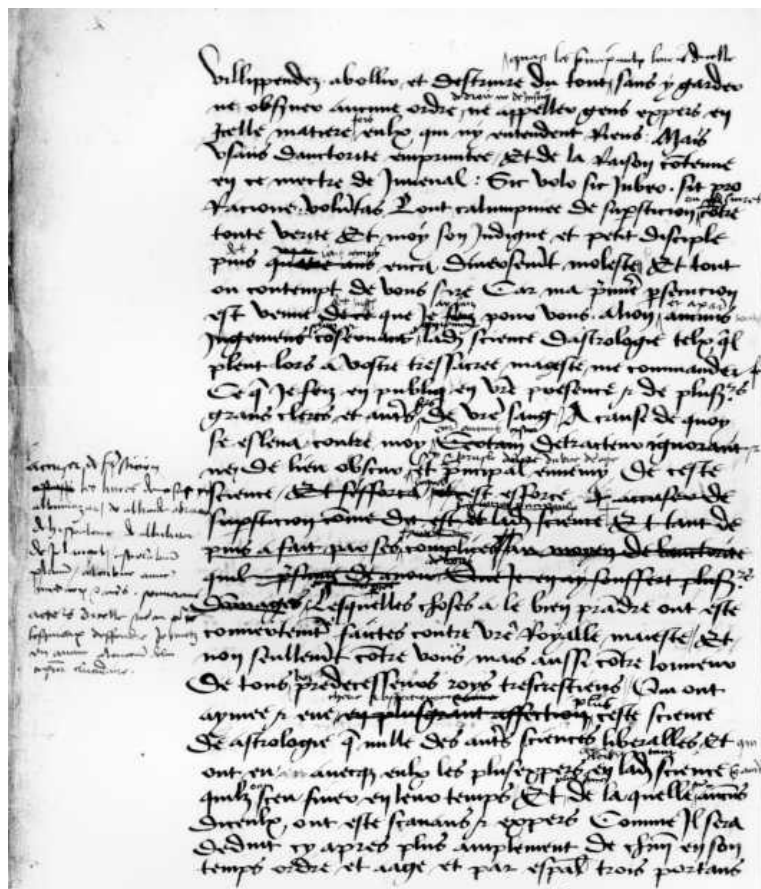
# Примеры перекрещивания структур



## ■ Лингвистическая / геометрическая

- «наложение» строк в конце абзацев
- перенос слов между строками и страницами
- нестандартная сегментация слов
  - lirois, idegnastesuenir
  - es paule, en mena

# Примеры перекрещивания структур



- Геометрическая / генетическая
  - исправленный текст расположен между строками или на полях
  - рукопись содержит несколько «слоев» исправлений

# Совместимость с лингвистической разметкой

- Необходимо обеспечить
  - возможность автоматической сегментации слов
  - возможность автоматической разметки предложений
  - разграничение «плоскостей» текста
    - основной текст / варианты / сноски и примечания
    - языки (если есть разные)

# Совместимость с лингвистической разметкой

- Классы элементов TEI по отношению к лингвистической иерархии
  - внутрисловные (с, g, am, ex...)
  - = 1 слово (num?, abbr?...)
  - одно или несколько слов в предложении
  - не менее одного предложения (div, p, head...)
  - вне структуры (note, foreign...)
  - плавающие (большинство элементов филологической разметки)
    - могут иметь «типичное» положение (например, внутри предложения)

# Пример спецификации разметки для корпуса текстов

- XML - BFM [http://bfm.ens-lyon.fr/IMG/pdf/Manuel\\_Encodage\\_TEI.pdf](http://bfm.ens-lyon.fr/IMG/pdf/Manuel_Encodage_TEI.pdf)
  - предварительная разметка слов, содержащих исправления букв или переносы: `<w>...</w>`
  - предварительная разметка сокращений и чисел, оформленных точкой
    - `<num>.i.</num>`, `<abbr>s.</abbr>`
  - отказ от использования элемента `<l>` в стихах:
    - `<lg> / <l> --> <ab> / <lg>` (противоречит стандарту TEI)
  - предварительная разметка «плавающих» элементов, охватывающих более 1 предложения
    - `<supplied rend="multi_s">`

# Пример спецификации разметки для корпуса текстов

## ■ XML - TXM

<http://sourceforge.net/apps/mediawiki/txm/index.php?title=Xml-txm-tei>

- расширение TEI
- все слова размечены как `<w>` с `@xml:id`
  - разметка (токенизация) автоматическая
  - возможность удаленной разметки со ссылкой на идентификатор слова
- каждый элемент `<w>` содержит
  - 1+ `<txm:form>`
  - 0+ `<txm:ana>`



# Спасибо!