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

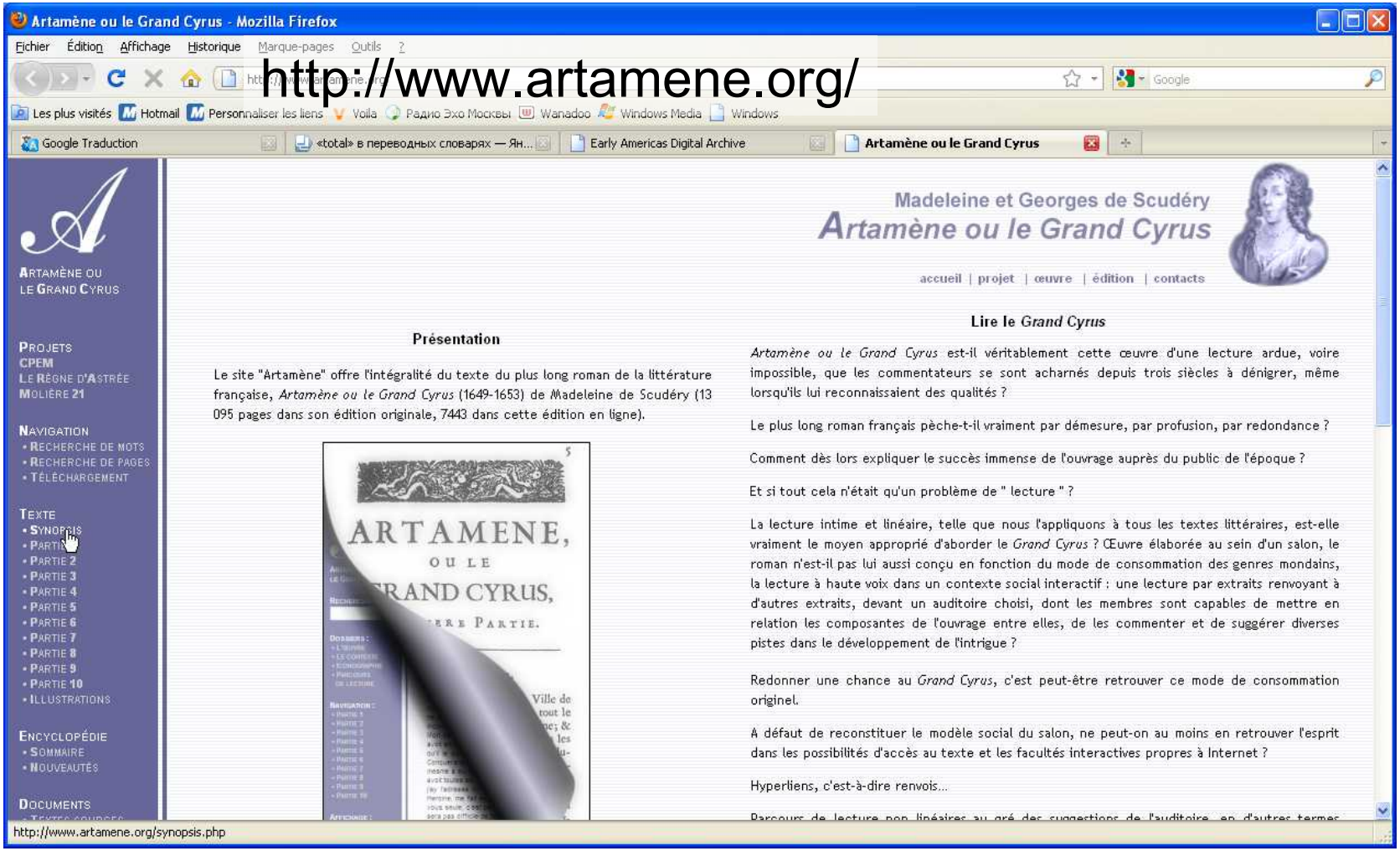
- Лекции Лу Бернарда (Казань, 2008)
  - Краткое введение в XML
  - Краткое введение в TEI
  - Основные структурные элементы TEI
    - http://tei.oucs.ox.ac.uk/Oxford/2008-08-kazan/

Проекты, использующие TEI

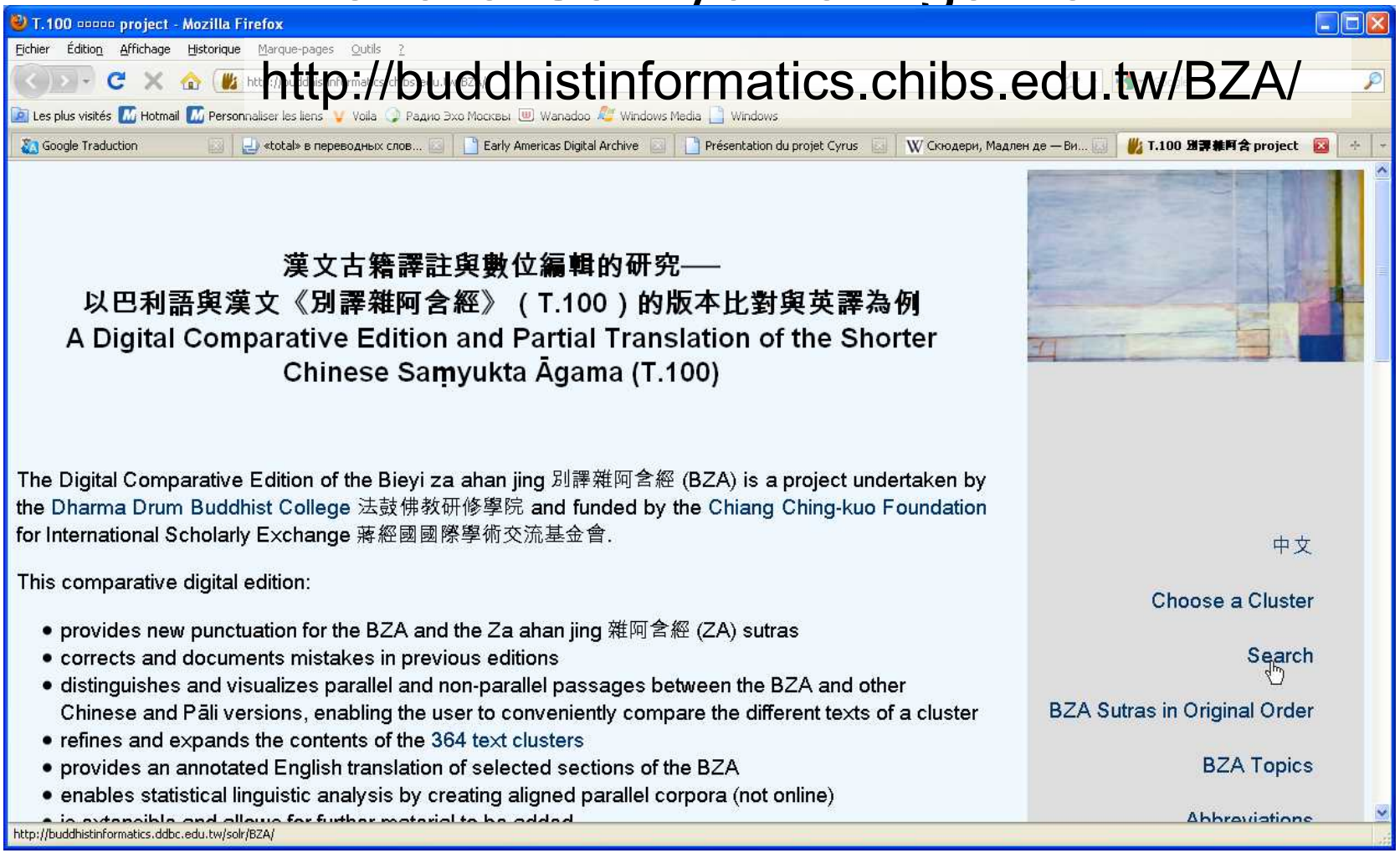
Early Americas Digital Archive

Women Writers Project

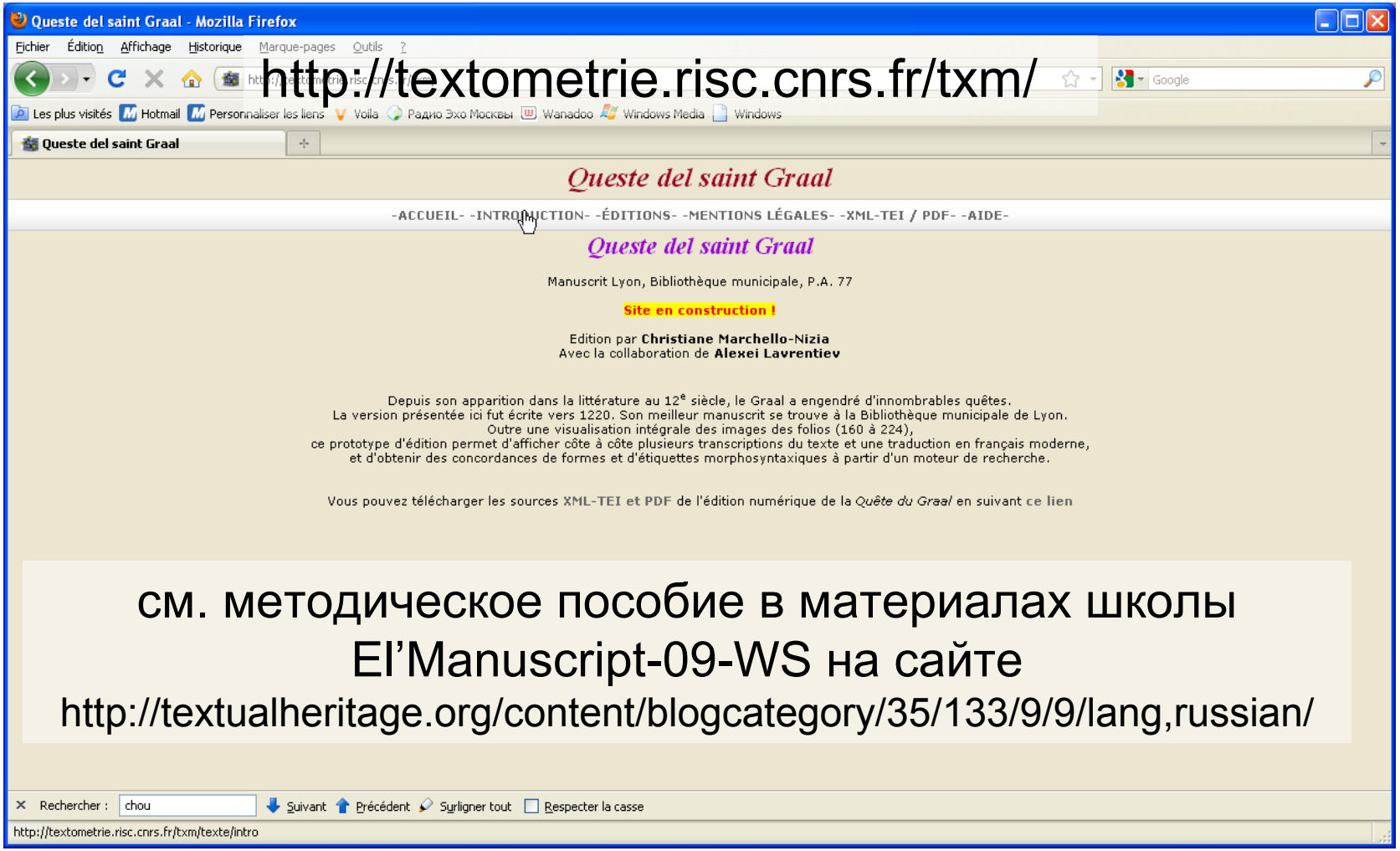
«Артамен, или Великий Кир »



Параллельное издание буддистского текста Samyukta Agama



Проект *Graal*

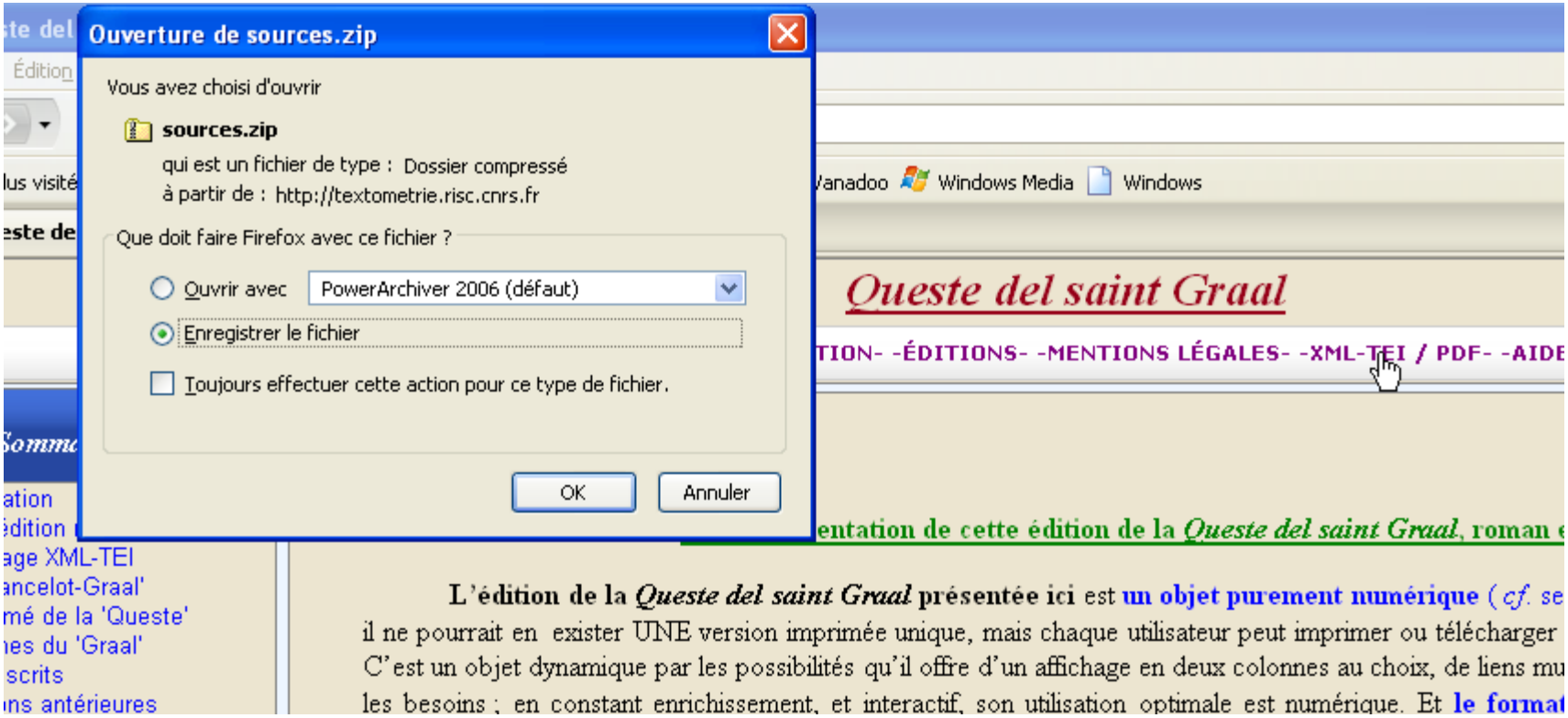


http://textometrie.risc.cnrs.fr/txm/

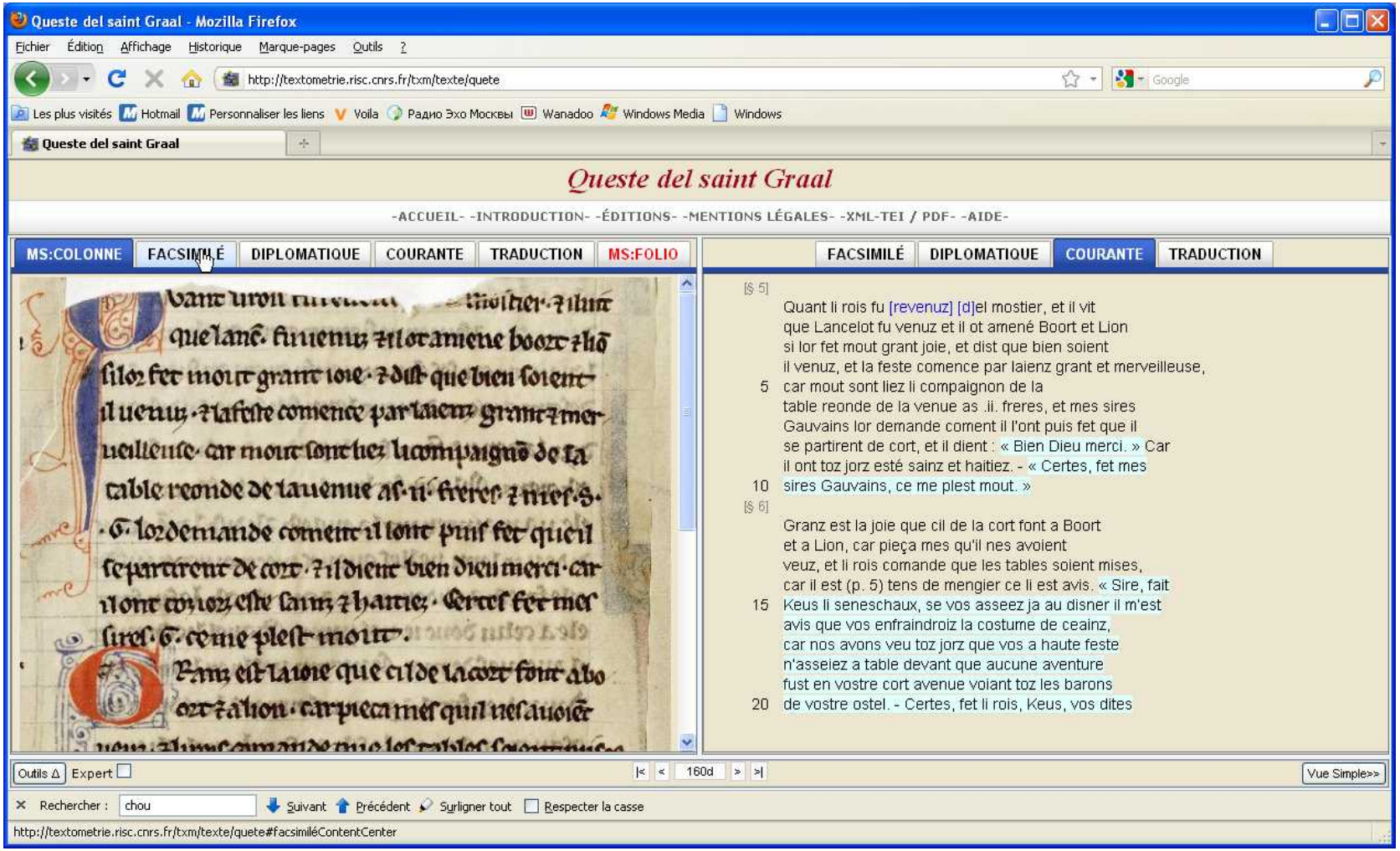
см. методическое пособие в материалах школы El'Manuscript-09-WS на сайте
http://textualheritage.org/content/blogcategory/35/133/9/9/lang,russian/

Проект *Graal*

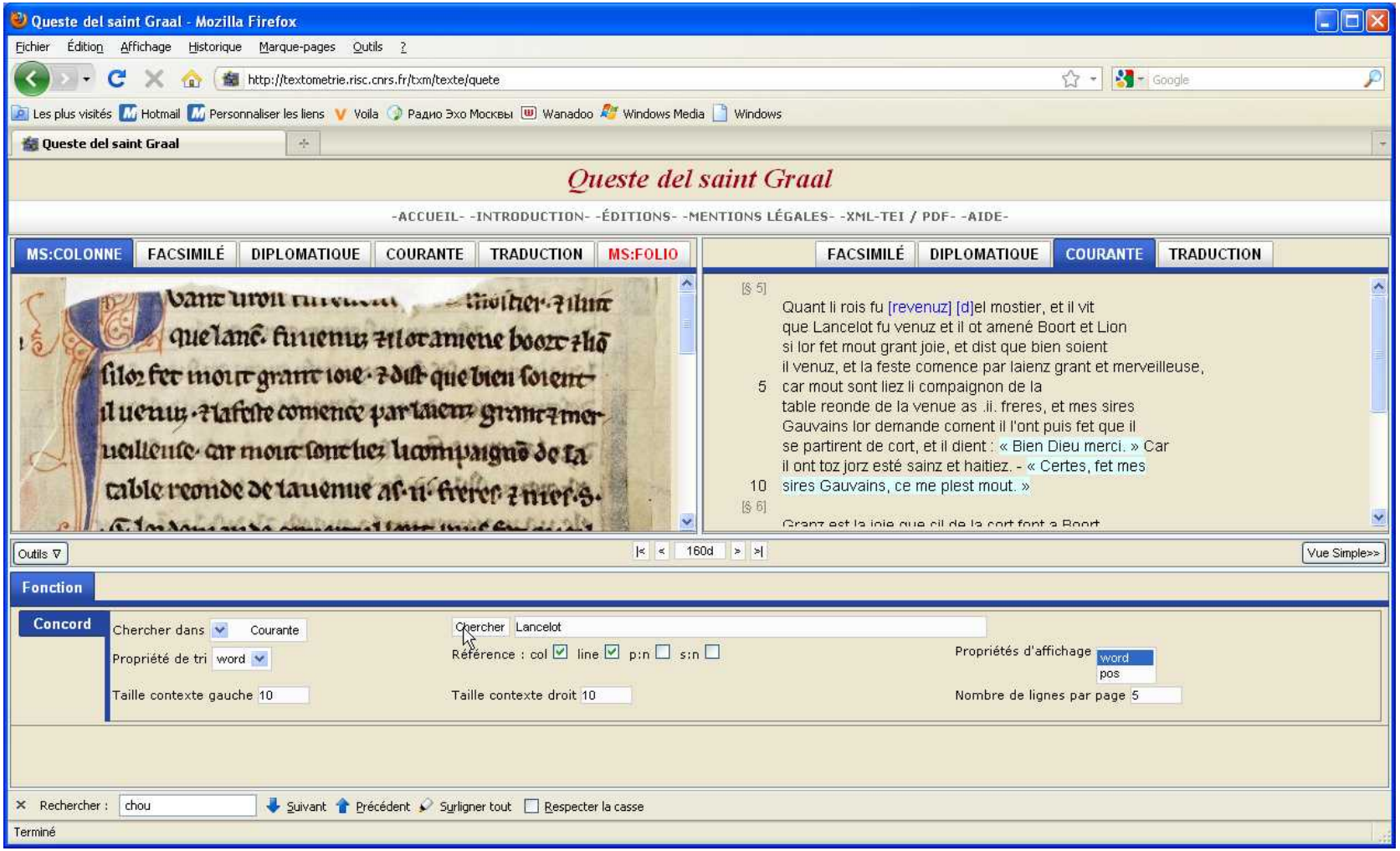
Проект *Graal*



Проект *Graal*



Проект *Graal*



Проект *Graal*

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
    - База средневекового французского http://txm.bfm-corpus.org/bfm/...
  - предлагают экспорт в формате TEI
    - Frantext http://www.frantext.fr/...

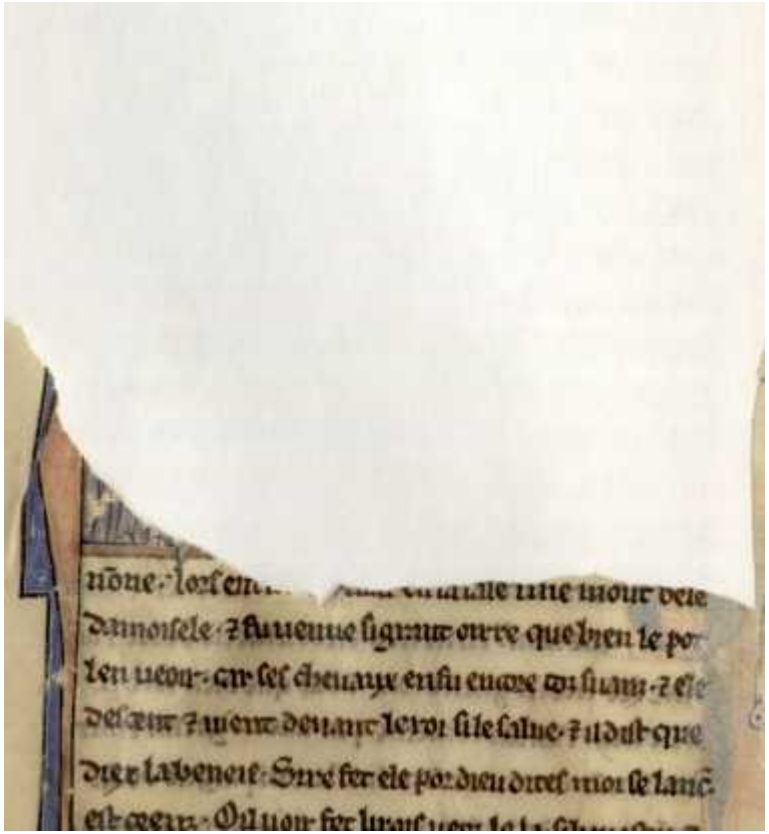
Стратегия разметки TEI

- 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

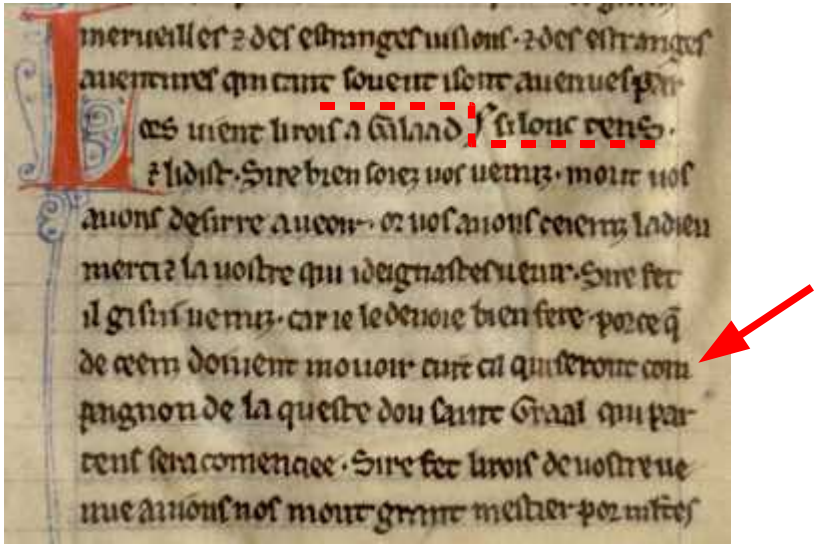
- С помощью ODD
  - Выбрать необходимые модули
  - Удалить ненужные элементы
  - Обосновать принятые решения
- При необходимости обратиться за помощью
  - TEI-L@LISTSERV.BROWN.EDU
- Собрать метаданные
  - teiHeader
- Подготовить стилевую таблицу CSS для редактирования

Примеры перекрещивания структур



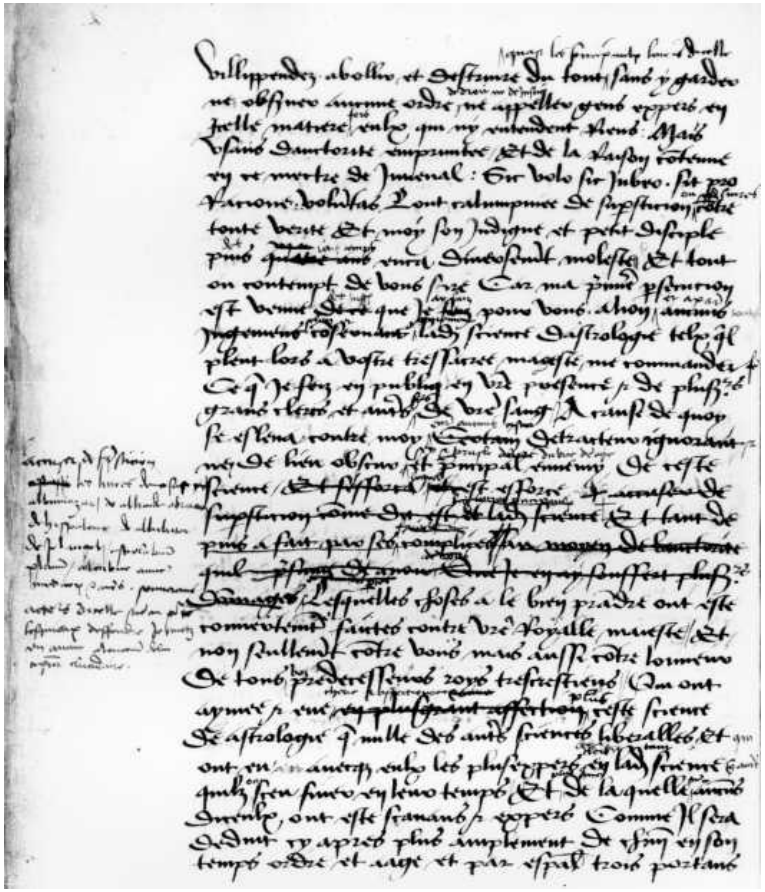
- Лингвистическая / редакторская
  - <supplied>A la veille de la Pentecoste […] tables a heure de</supplied> nonne lors en<supplied>tra a cheval</supplied> <unclear>en la s</unclear>ale

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

Ms. BnF fr. 1357, f° 2v, Simon de Phares, *Recueil des astrologues*

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 (c, 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
  - 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!