Оцифровка

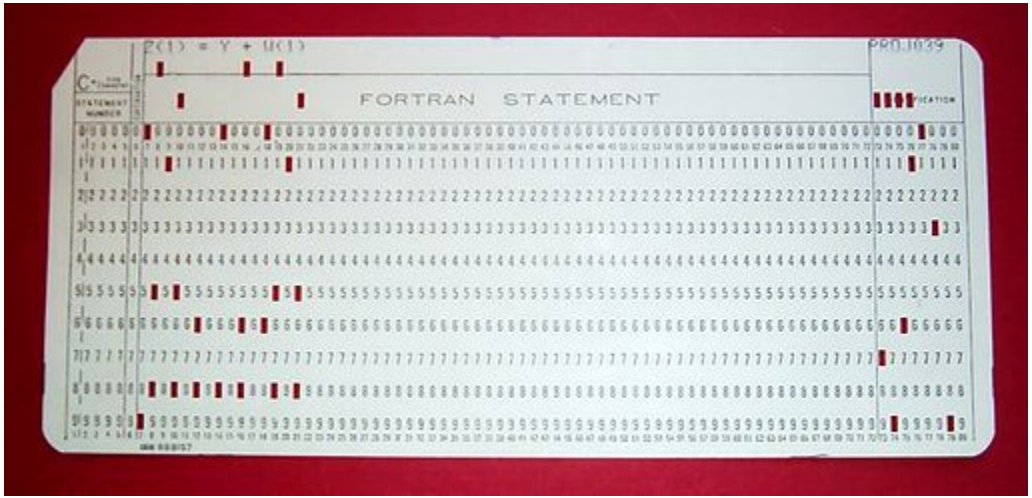
Лекция № 2

План

- Что такое оцифровка и почему её недостаточно
- Оцифровка как проблема и как тренд DH
- Концептуальная сторона: текст vs. документ
- От традиционной текстологии к цифровой
- Digital surrogate: неполнота цифровой копии
- Разметка и стандарт TEI
- Семантическое издание и новые практики

Оцифровка — это кодирование

- Оцифровка = перевод информации в машиночитаемую форму
- Не только «картинка», но и структура, метаданные, связи
- Результат — не файл, а представление объекта в новой семиотической системе



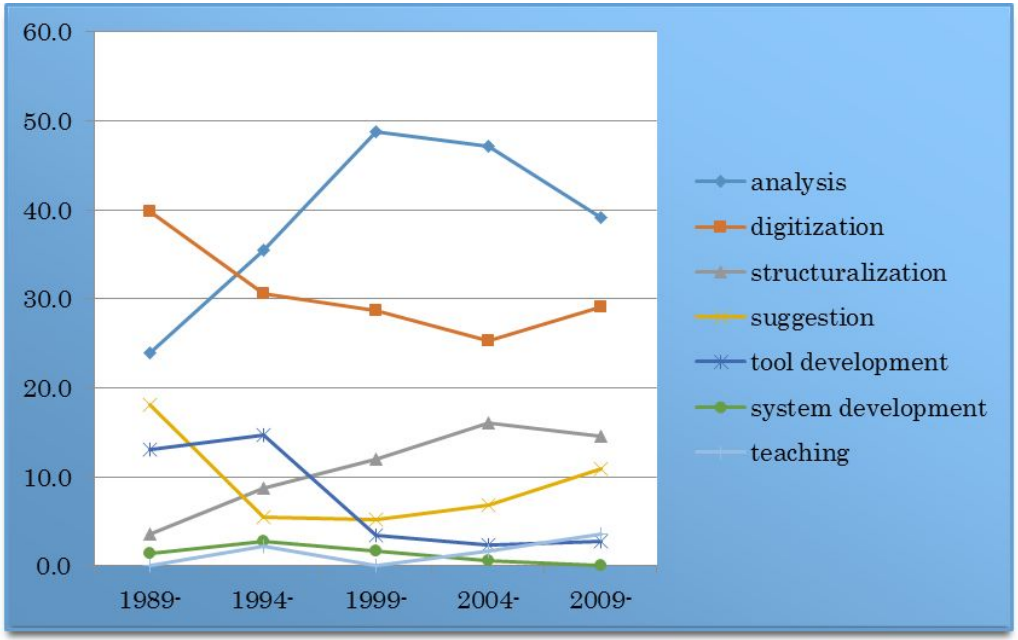
Отсканировать недостаточно

- Скан фиксирует внешний вид, но не смысл
- Без разметки текст остаётся «невидимым» для машины
- Нужны транскрипция, структура, аннотации



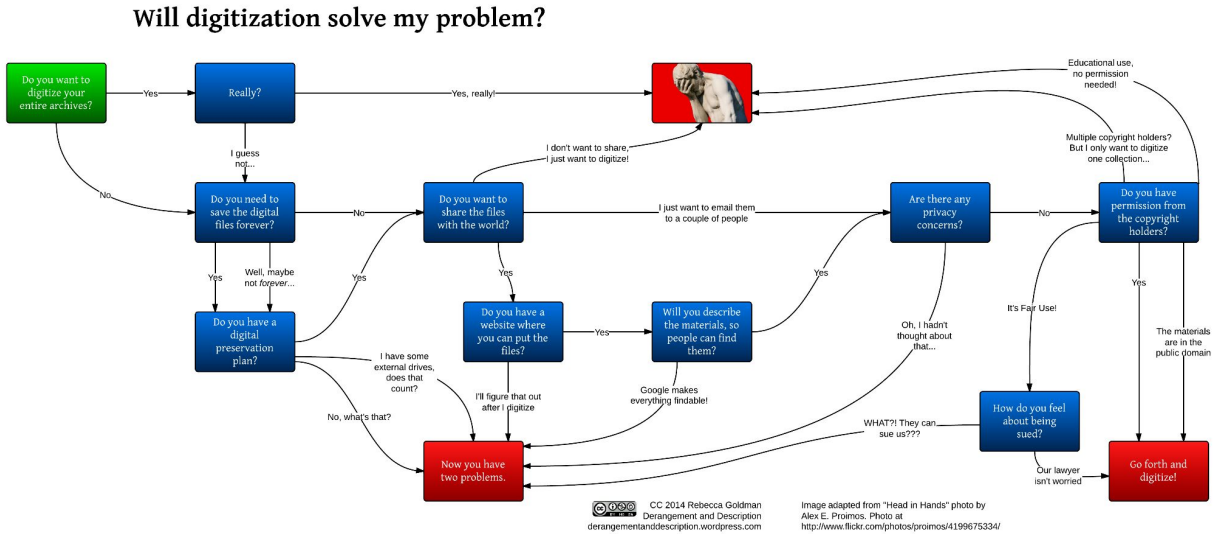
Оцифровка в трендах DH

- Цифровые архивы и коллекции
- Агрегаторы born-digital материалов
- Специальные коллекции (например, The September 11 Digital Archive)



Оцифровка — это проблема

- Что именно сохранять?
- Как отделить важное от второстепенного?
- Как передать свойства, не выразимые буквами?



Оцифровка — это сложно

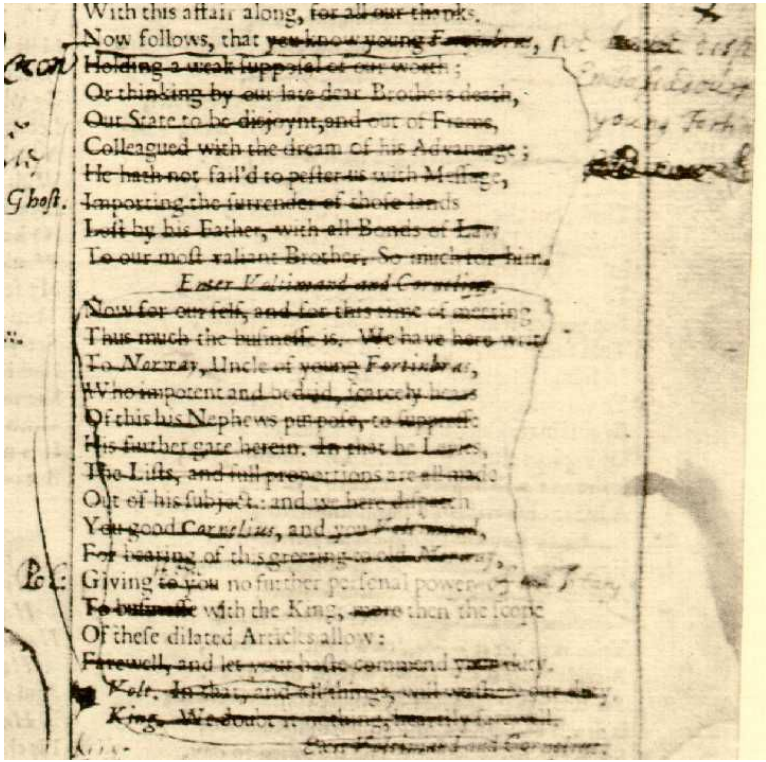
- Материальный носитель, почерк, правки, повреждения
- Многослойность документа
- Разные цели у лингвистов, историков, литературоведов

Контекст оцифровки

- Неизбежна: переход человечества на цифровые носители
- Сохранение культурных артефактов
- Удешевление копирования и доступа к информации
- Электронные публикации могут включать больше документов
- Новые социальные практики (shared knowledge)

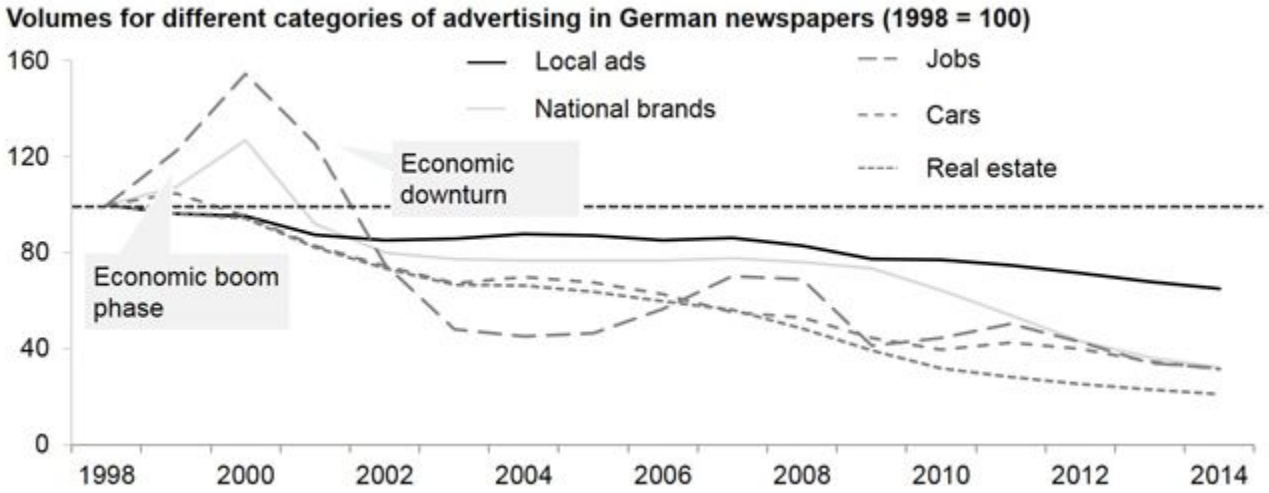
Текст vs. документ

- Дуализм: текст как документ и текст как произведение
- Цифровое издание лучше представляет и то, и другое
- Документ — материальный носитель; текст — конструкция читателя



Как оцифровывать?

- Почему актуален вопрос «как оцифровывать?»
- Важными могут оказаться неожиданные аспекты
- Если сохранять только важное — как решить, что важно?





Что важно тут?

Традиционная текстология

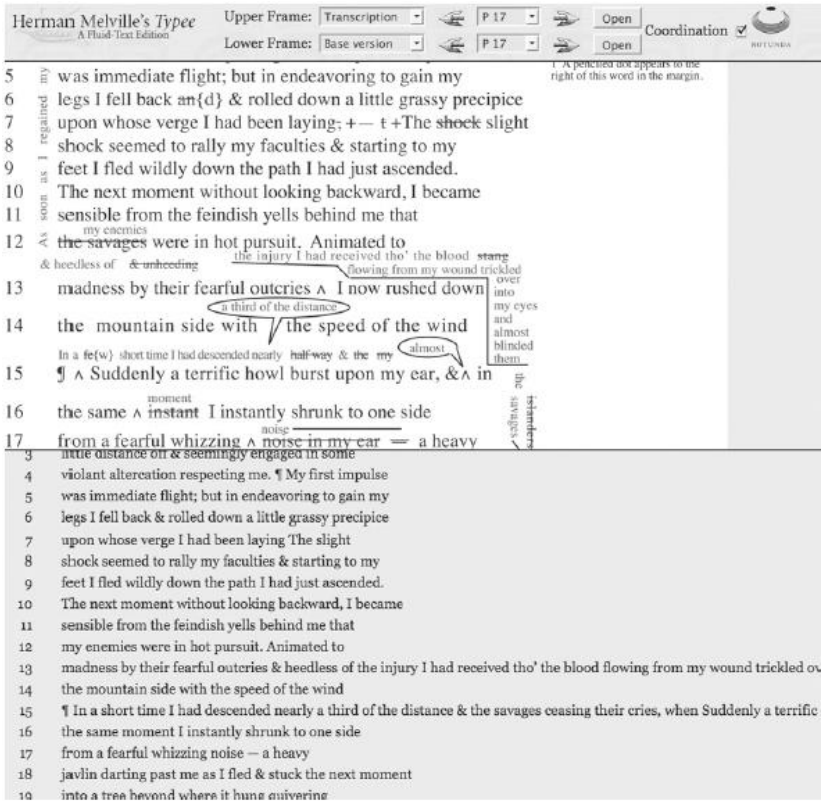
- История текста как единое целое: слои и контекст
- Литературная правка и внелитературные факторы (цензура)
- Отделение текста (сознательные исправления) от не-текста (ошибки, рисунки)



Рис. 3. Типы изданий древних текстов и рукописей

Дипломатические и критические издания

- Критическое издание: реконструкция авторской воли
- Дипломатическое издание: точная передача документа
- Разные типы изданий — разные задачи



Текстологические задачи печатной эпохи

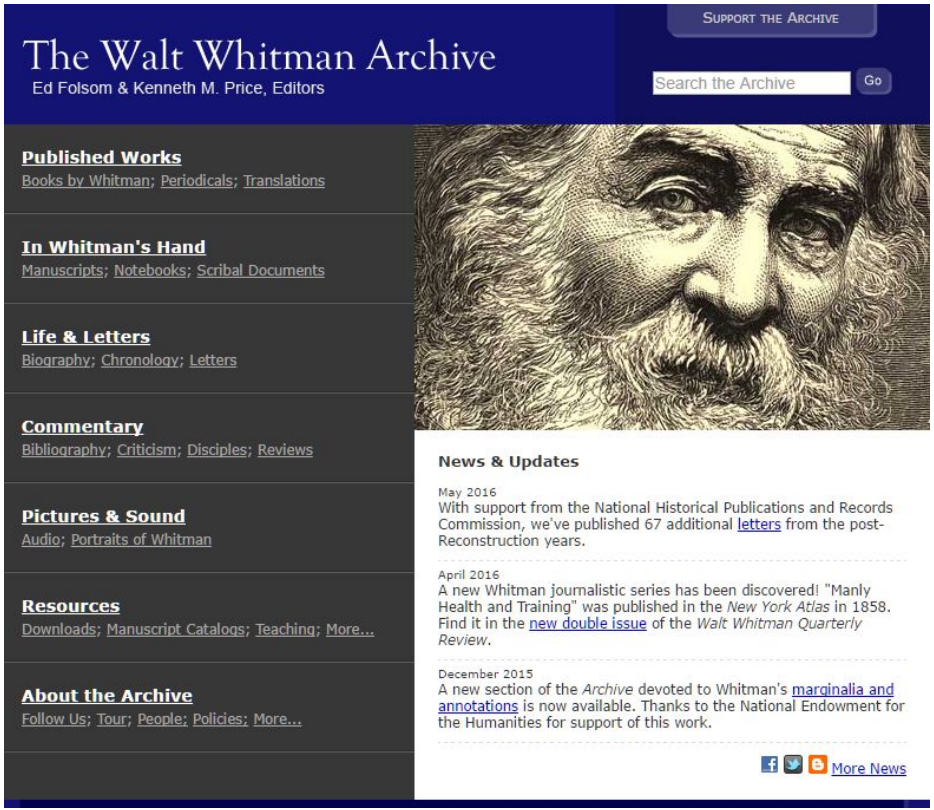
- Аван-текст: черновик / беловик / копия
- Первая публикация
- Прижизненные издания
- Разночтения и редакции
- Принцип последней авторской воли

Цифровая эпоха

- «Текст следует рассматривать принципиально как функцию документа. Ведь в конечном счёте… у нас есть документы, и только документы» (Gabler 1995)
- Уход от реконструкции воли автора к точной передаче текста документа
- Экономические причины: удешевление дипломатических изданий и публикации больших архивов
- Технологические причины: электронная публикация вовлекает читателя и облегчает восприятие многослойного документа

Академический подход к цифровому изданию

- The Walt Whitman Archive: «подход архива — устанавливать документальные тексты, а не реконструировать то, что считается авторски задуманными текстами»
- Внимание к историческим и социальным аспектам текстов
- Каждый шаг оцифровки — суждение редактора



Вопросы цифровой текстологии

- Вместо «что является авторской волей?» —
- «что является значимым в документе?»
- «что является валидным текстом документа?»

Digital surrogate

- «Цифровой суррогат» — цифровая копия аналоговой записи
- Общая черта: неполнота (избирательность) представления информации исходного источника
- Причины создания: защита оригинала, доступность, более читаемая версия
- Риск: подмена источника его неполной копией

Что необходимо сохранить?

- Текст:
  - лингвистическое содержание
  - особенности начертания (часть свойств передаётся шрифтом)
- Материальный носитель и его особенности
- Метаданные, исправления, конвертируемость в современный вид

Разметка

- Специальные символы, делающие обработку текста автоматической
- Указания для компьютера
- Сделать явным для машины то, что подразумевается читателем

```html
<!DOCTYPE html>
<html>
    <head>
        <meta charset="utf-8">
        <title>Язык разметки</title>
    </head>
    <body>
        <h1>Язык разметки<h1>
        <p><strong>Язык разметки</strong> в компьютерной терминологии — набор
            символов или последовательностей, вставляемых в текст для передачи
            информации о его выводе или строении. Принадлежит классу
            компьютерных языков.</p>
        <h2>Структура языка</h2>
        <p>Текстовый документ, написанный с использованием языка разметки,
            содержит не только сам текст но и дополнительную информацию:</p>
        <ul>
            <li>указание на заголовки</li>
            <li><mark>выделения</mark></li>
            <li>списки</li>
            <li>и т. д.</li>
        <p>В более сложных случаях язык разметки позволяет вставлять в
            документ <i>интерактивные элементы</i> и содержание других
            документов.</p>
    </body>
</html>
```

Язык разметки — HTML Academy

Структура языка

Текстовый документ, написанный с использованием языка разметки, содержит не только сам текст но и дополнительную информацию:

- указание на заголовки
- выделения
- списки
- и т. д.

В более сложных случаях язык разметки позволяет вставлять в документ *интерактивные элементы* и содержание других документов.

Стандарт TEI

- Text Encoding Initiative — стандарт электронной публикации
- Сообщество, вырабатывающее стандарт (консорциум: 71 институт, 190 участников)
- Разметка на основе XML
- Смысловые модули: метаданные, структура, критика, лингвистика



Семантическое издание

- С помощью разметки выделяются не только структурные, но и семантические элементы
- Текст становится понятным компьютеру
- Элементы увязываются между текстами
- Это позволяет говорить о семантическом издании

- «Возможно, нам стоит просто перестать пытаться отобразить цифровые издания как печатные и вместо этого признать, что мы производим объект иного типа, который можно назвать документальным цифровым изданием» (Pierazzo)
- Такое издание включает три компонента: источник, результат и инструменты
- Оцифровка — не копирование, а перевод между семиотическими системами
- Стандартизация → сохранение → агрегация