Enabling complex analysis of large-scale digital collections: humanities research, high-performance computing, and transforming access to British Library digital collections

Обеспечение комплексного анализа крупных цифровых коллекций:
гуманитарные исследования, высокопроизводительные вычисления и изменения доступа к цифровым коллекциям Британской библиотеки

A pilot project at University College London

Параметры исследования:

1. Вопрос:

Как ученые-гуманитарии могут наиболее эффективно пользоваться крупномасштабными цифровым коллекциями, доступными в учреждениях культуры?

2. Участники проекта:

3. Цифровой архив:

Крупномасштабная цифровая коллекция Британской библиотеки с открытым доступом. 60 000 оцифрованных книг.

В этом проекте были уточнены технические и процедурные барьеры, возникающие, когда исследователи-гуманитарии пытаются использовать электронные инструменты культурных институций для решения своих исследовательских задач.

Методология:

- 4 месяца
- 4 исследователя / исследования
- Разработчики программного обеспечения обрабатывают исследовательский вопрос, приводя его в форму запросов, ответы на которые можно получить с помощью электронной обработки цифрового архива и создают инструменты для получения нужных результатов.
- Наблюдение за процессом, выявление «узких мест»
- Интервью с участниками на стадии работы над проектом

Проекты 1 и 2:

1. Задача:

Поиск слов и окружающего их контекста (примеры: professor, higher education)

2. Что нужно на выходе:

Страница окружающая искомое слово. Массив из всех результатов.

2. Что дальше:

Close reading исследователем.

Проект 3:

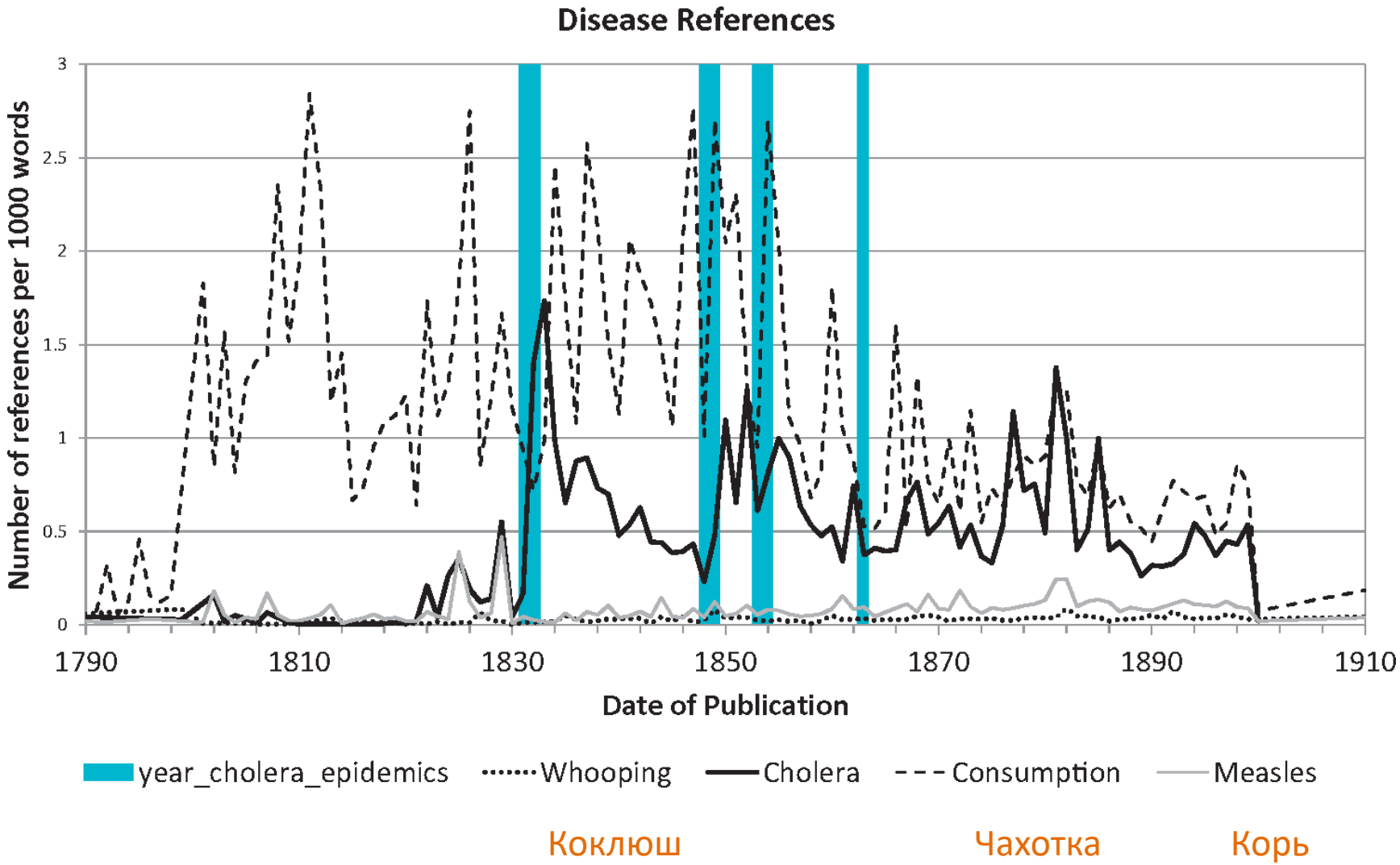
ИСТОРИЯ МЕДИЦИНЫ

Исследовательский вопрос:

Как частота упоминания тех или иных заболеваний в литературе соотносится с известными эпидемиями XIX века?

Можем ли мы увидеть какую-либо корреляцию между возникновением инфекционных заболеваний в обществе и упоминанием этих болезней в художественной и научно-популярной литературе?

Проект 3: диаграмма



Проект 4:

ИСТОРИЯ ИЛЛЮСТРАЦИИ

Изменения в технологии печати, произошедшие в период с 1750 по 1850 год, привели к появлению нескольких новых типов иллюстраций, а также к новым, неожиданным способам верстки (то есть к новым типам соотношения и соположения картинки и текста).

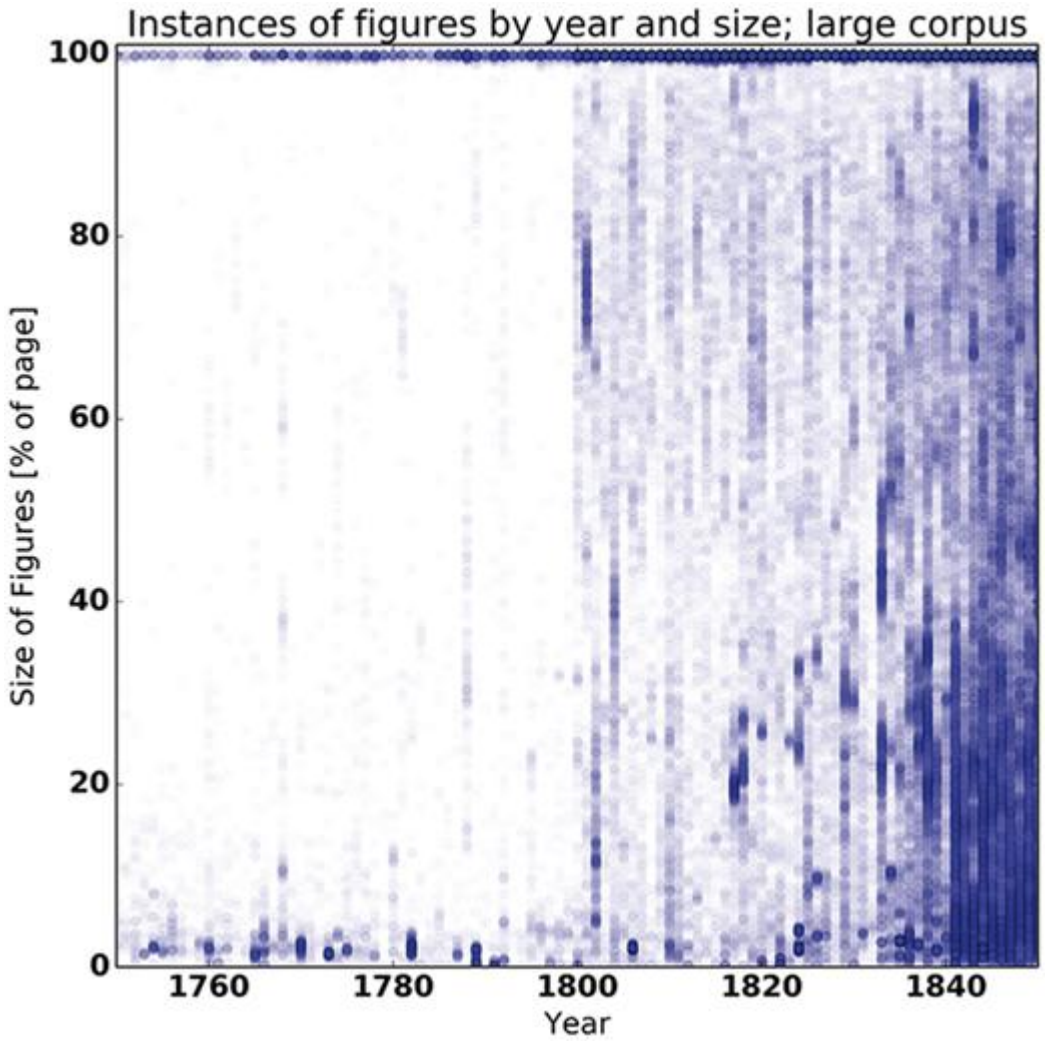
Исследовательские вопросы:

Как изменения в методах печати и в размерах изображений отражались на разных типах («жанрах») иллюстраций с течением времени?

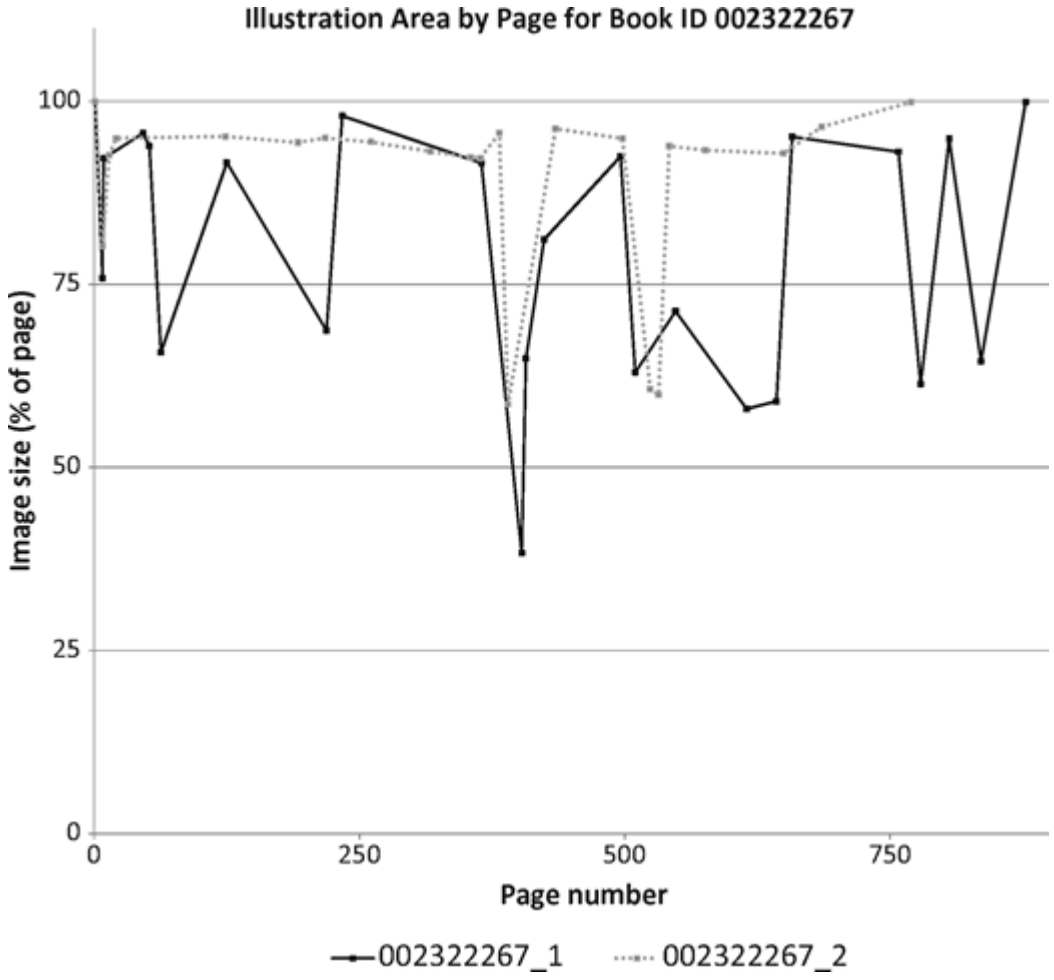
Можно ли с помощью количественныех данных понять, как изменения в стилях иллюстрации меняли ее значение?

Какие выводы можно сделать при помощи dh-технологий и цифровых коллекций? Отличаются ли они от выводов, полученных традиционными методами на материале небольших частных коллекций?

Проект 4: диаграмма 1



Проект 4: диаграмма 2



Выводы: препятствия

Барьеры для широкого использования цифровых коллекций и связанных с ними вычислительных подходов исследователями-гуманитариями:

- Разобщенность сообщества, ресурсов и инструментария;
- Недостаточность взаимодействия и координации участников;
- Разнородность цифровых коллекций, отсутствие гетерогенности в инфраструктуре;
- Недостаток технических навыков.

ВЫХОД

Создание единых веб-интерфейсов или общей инфраструктуры для цифровых коллекций

Выводы: решения

Практические рекомендации, выявленные в ходе работы:

- необходимо создать большое количество стандартных инструментов работы с данными (счетчики книг и слов в год, слов и страниц на книгу и т. д.) для нормализации результатов и предоставления быстрого доступа к статистике;
- необходимо документировать решения, найденные в процессе работы с данными и метаданными;
- Необходимо выдавать на выходе четкую, поддающуюся анализу и описанию информацияю, которые ученые могли бы трактовать и использовать (одновременно стоит учитывать риски связанные со стандартизированием информации)

ВЫХОД

1. Учреждения культуры должны **инвестировать в разработку программного обеспечения** для открытых крупномасштабных цифровых коллекций, и таким образом содействовать исследованиям в области искусств, гуманитарных, социальных и исторических наук.
2. А также **инвестировать в обучение библиотечного персонала**, который мог бы обрабатывать первичные запросы исследователей-гуманитариев, поддерживать работу баз данных, документировать полученные результаты и хранить производные данные.

Критика исследования:

- Проблема, которой занимается исследование видна невооруженным взглядом. То, что цифровые коллекции выкладываются без прилагающихся инструментов анализа, – очевидно. Что это серьезный барьер для ученых из гуманитарной сферы – также довольно прозрачный факт.
- Три из четырех сделанных в ходе исследования проекта не представляют особого интереса. В первых двух случаях достаточно использовать разработки Корпуса, в третьем (про болезни) плохо поставлен исследовательский вопрос – ответ очевиден и без исследования.
- Выводы в статье, тем не менее, довольно релевантны как для учреждений культуры, так и для исследователей-гуманитариев. Однако, на практике проблема создания инфраструктуры и связей между участниками остается открытой.