Инна Кижнер

КОЛЛЕКЦИИ КАК ДАННЫЕ: ГОТОВЫ ЛИ МЫ К НАУЧНЫМ ИССЛЕДОВАНИЯМ С ПОМОЩЬЮ ЦИФРОВЫХ КОЛЛЕКЦИЙ КУЛЬТУРНО-ЗНАЧИМЫХ ДАННЫХ?

Задача настоящей работы — показать ограничения, которые существуют на пути к использованию цифровых коллекций культурных учреждений как источника данных. Цифровые коллекции предстают инструментом познания. Они удивительно редко используются как данные. У этого обстоятельства есть ряд обсуждаемых в статье причин. Кроме того, существенно, что объекты в цифровых коллекциях распределены неравномерно и подобраны субъективно. Все эти особенности влияют на результаты научного анализа и их необходимо учитывать в ходе цифрового исследования.

Ключевые слова: цифровые коллекции, данные, цифровое неравенство, музейные коллекции

Введение

Еще недавно цифровые коллекции библиотек, архивов и музеев воспринимались только как инструмент поиска, ускоренного доступа к первичным источникам и чтения с экрана. Недавний перелом в отношении к цифровым коллекциям ведет к переосмыслению роли собраний цифровых копий артефактов (текстов, изображений и аудиофайлов). Представление о цифровых коллекциях как об источнике данных — «коллекции как данные» – делает собрания библиотек музеев и архивов одним из немногих мест, где большие датасеты могут использоваться как эвристический инструмент. При

Инна Александровна Кижнер
Университет Хайфы
inna.kizhner@gmail.com

DOI: 10.31860/cgi-2024-1-45-62

этом возникает возможность пользоваться коллекциями в рамках разных подходов, используя разные инструменты. Использование коллекций как данных в разных проектах и разными научными коллективами полезно не только для верификации предшествующих результатов, распространения полученного знания и планирования дальнейших исследований, но и для расширения границ применимости инструментов и интерпретации результатов [Borgman 2012, Sköld et al. 2023]. Это выгодно отличает коллекции библиотек, архивов и музеев от закрытых личных или проектных коллекций с ограниченным использованием. Можно ли приблизиться к объективному знанию, используя цифровые коллекции музеев, архивов и библиотек как данные? Какие препятствия существуют на этом пути? Должны ли ограничения, присущие цифровым коллекциям, быть эксплицированы и объяснены исследователю?

Возможность неоднократно обращаться к данным существует и для открытых репозиториев в академических учреждениях, таких как университеты и научно-исследовательские институты. Репозитории научных организаций или проектные репозитории выполняют функцию агрегирования и курирования данных в отдельной научной области, такой как археология или литературоведение. Однако при анализе разных способов классификации, кластеризации и представления исторических объектов и документов имеет смысл рассмотреть библиотеки, музеи и архивы. Такие организации, предназначены для долгосрочного хранения и обработки информации. Они получают постоянное финансирование, а некоторые из них существуют многие десятки и сотни лет. Там работают специально обученные люди, квалификация которых позволяет объединить разные данные в коллекции и разработать способы анализа. Несомненно личные, проектные и институциональные цифровые коллекции обладают дополнительными преимуществами навыков обработки данных в узкой области или высокой квалификации инициатора и организатора репозитория, но они подвержены разным рискам, связанным с прекращением финансирования или сменой интересов организатора.

Задача настоящей работы — показать ограничения, которые существуют на пути к использованию цифровых коллекций культурных учреждений как источника данных. Интересно, что анализ препятствий показывает возможности цифровых коллекций как инструмента познания. При этом анализ данных, полученных при изучении самого предмета, осложняется тем контекстом, который в цифровой коллекции определяет место предмета и документа.

Имеют значение не только обстоятельства поступления предмета в коллекцию, но и обстоятельства, при которых коллекции были сформированы, а предметы и документы классифицированы и каталогизированы. Принципы создания физических и цифровых коллекций определяются политическими, экономическими и социальными причинами, которые влияют на академические традиции и научные интересы коллекционеров.

Цифровые коллекции культурно-значимых данных вошли в нашу жизнь вместе с цифровыми инструментами анализа. Преимущества анализа больших датасетов, такие как возможность обобщений и воспроизводимость исследований, поддерживаются значительно большим количеством примеров чем анализ, основанный на небольшом количестве примеров в случае медленного чтения. Количественный и статистический анализ, который приводит к кластеризации данных, выигрывает от возможности создать стратифицированную выборку. Инструменты познания, ранее привязанные к воображению и риторическим приемам, при наличии больших датасетов и цифровых коллекций становятся инструментами, которые приводят к созданию универсальных теорий, - разумеется, в рамках своих возможностей [Piper 2020].

Возможность воспроизводимых исследований с использованием коллекций как данных часто связывают с возможностью публиковать данные с лицензиями, которые разрешают неограниченное использование (см., например, [Wallace and Euler 2020, Vezina et al. 2022, Candela et al. 2023]). Другие необходимые условия могут включать наличие интерфейса прикладного программирования (API), который обеспечивает возможность переноса большого количества данных на машины научного коллектива и дополнительные возможности анализа. Еще одно важное условие - предоставление ссылок (возможности цитирования и повторного обращения к наборам данных), таких например как DOI [Маслинский 2022, Candela 2023]. Последнее обстоятельство обеспечивает валидацию результатов исследований, поскольку при смене способов публикации или при переходе на другую платформу, датасет все равно будет обнаружен. При этом учреждение, которое публикует данные несет ответственность за предоставление доступа и возможность обнаружения датасета.

Действительно, некоторые из этих условий соблюдаются ограниченным количеством библиотек, архивов и музеев. Некоторые из этих институций обладают значительным влиянием и обширными коллекциями. Для других учреждений важно использовать

открытые лицензии для небольшой части коллекции. В июле 2023 года список библиотек, музеев, архивов и агрегаторов культурно-значимых данных, которые используют открытые лицензии включал около 1650 учреждений из 64 стран мира [Wallace and McCarthy 2023]. Географическое распределение стран отражает (разумеется, с некоторыми исключениями) распределение стран с высокими и средними доходами на душу населения. Ряд музеев и библиотек действительно использует интерфейс прикладного программирования (API) для предоставления доступа к своим данным[1]. Подобные датасеты, в частности, используются для обучения алгоритмов компьютерного зрения для классификации и генерации произведений искусства (см., например, [Conde and Kurgutlu 2021, Castellano and Vessio 2021, Rei et al. 2023]). Эти источники данных включают изображения в разных форматах и жанрах и сопровождаются текстовыми метаданными, которые характеризуют период, формат и культуру, в рамках которой создано изображение. Это важно, потому что дает возможность создания обучающей выборки для последующей работы алгоритмов и решения задач классификации и генерации изображений.

Однако несмотря на соблюдение ряда условий некоторыми библиотеками, музеями и архивами, удивляет редкое использование цифровых коллекций культурных учреждений как источника данных. Так например, из тринадцати статей журнала “Journal of Computational Literary Studies” 2022 год только три статьи ссылаются на цифровые коллекции научных коллективов, институций и фондов (Project Gutenberg[2], Perseus Digital Library[3], European Literary Text Collection[4]). Ни одна из статей не ссылается на цифровые коллекции библиотек и архивов. В пятидесяти статьях сборника расширенных тезисов конференции Digital Humanities 2022[5] только шесть статей упоминают такие коллекции в качестве источника данных. В чем же причины предпочтения, которое ученые оказывают личным и проектным коллекциям. Почему обширные цифровые коллекции, опубликованные с открытыми лицензиями, а иногда и с ключом API, не используются исследователями?

Почему коллекции не используются как данные? Причины создания авторских или проектных корпусов

Среди препятствий, который осложняют использование цифровых коллекций как источника данных, можно выделить следующие обстоятельства:

- Непоследовательность в индексации данных даже тогда, когда в культурном учреждении есть внутренний стандарт метаданных;

- Отсутствие данных, опубликованных в форматах, которые позволяют проводить машинный анализ;

- Неоднозначность и нечеткость культурно-значимых данных, уникальные описания объектов, которые включают точку зрения, характерную для исследователя, который изучал этот объект в прошлом;

- Пропуски в значениях метаданных, для значительной части объектов некоторые поля пустуют;

- Вопрос исследования требует добавочного поля или нескольких полей;

- Коллекция музея, архива или библиотеки недостаточно репрезентативна и сбалансирована;

- Коллекция музея, архива или библиотеки не включает художественные произведения того канона, который интересует исследователя.

Отсутствие стандартизованных и нормализованных данных

Отсутствие единого стандарта данных на международном, национальном и даже институциональном уровне приводит к тому, что сложно объединить данные, которые относятся к одному типу, периоду или географической локации. Это обстоятельство является важным препятствием на пути к количественной обработке и обобщениям. В первую очередь неравномерное распределение результатов анализа данных в коллекции может быть вызвано пропусками в данных, которые могут достигать больших значений, например, в поле метаданных «Место издания книги» в цифровой коллекции Библиотеки Эдинбургского университета стоят пропуски в 70% случаев [Havens et al. 2022]. Помимо пропусков в полях метаданных большую проблему составляет разнообразие типов, жанров и форматов гуманитарных данных. Описание и документация такого разнообразия проводятся по-разному в зависимости от той академической традиции, к которой принадлежат сотрудники

библиотек, музеев и архивов. Даже в одном и том же учреждении формат записи даты, географической локации или решения принятые относительно жанра, техники или темы работы могут (и должны) отличаться для значений полей метаданных, а иногда и для полей метаданных в системе управления коллекциями, если система позволяет добавлять новые поля. Совместимость полей и/или значений полей метаданных, а также технологические решения, обеспечивающие совместимость данных цифровых коллекций, обсуждаются в информационных исследованиях, начиная с последних десятилетий 20 века (см., например, [Bearman 1995, Besser 2002, Doerr 2014, Tolfo et al. 2021]). Отдельной задачей совместимости данных является нормализация значений метаданных. Эта задача, в частности, основана на совместимости методов записи дат, географических локаций и персоналий. Например, в археологических коллекциях способы обозначения дат, и тем более записи текстовых выражений, обозначающих приблизительность датировки, значительно варьируются в зависимости от принятого метода или академической традиции [Binding and Tudhope 2023]. Несмотря на многочисленные попытки ввести словари, онтологии и объединенные списки для унифицированной записи культурно-значимых данных, такие как словари Гетти [Harping, 2013] или словари, принятые внутри одного учреждения[6], а также попытки перевода словарей Гетти на другие языки (см., например, [da Silva 2022]), проблема нормализации данных все еще остается нерешенной даже в академических культурах с широкими традициями обработки гуманитарных данных. Она остается особенно важной в контексте анализа культурно-значимых данных из коллекций Музейного фонда РФ, опубликованных в связи с созданием Государственного каталога Музейного фонда, который на момент написания статьи включает около 40 миллионов учетных записей [Глазунов, Орехов 2020, Костенко, Козлова 2021].

Одна из причин отсутствия стандарта метаданных или неохотного следования стандарту в случае, когда справочники и словари доступны, — неоднозначность и нечеткость культурно-значимых данных. Разнообразие точек зрения на объект или явление, разница в академических традициях и идеологической позиции приводит к тому, что данные цифровых коллекций представляют собой не только разные слои подходов к коллекциям при переходе от физической коллекции к цифровой [Mak 2014], но и разные варианты описаний объектов. При этом варианты описаний могут быть созданы в разные эпохи, разными людьми, при разных социальных обстоя-

тельствах [Parry 2007, Cameron 2010, Gnoli 2011, Zhitomirsky-Geffet 2019]. С одной стороны эффект “пограничного объекта” [Star and Griesemer 1989] осложняет создание инфраструктуры, в которой коллекции могут рассматриваться как данные и/или эвристический инструмент. С другой стороны нечеткость и неопределенность места в классификационной схеме присущи и естественно-научным объектам и тоже определены контекстом, социальными обстоятельствами и академическими традициями. Эти проблемы, вероятно, могут быть решены с помощью развернутых онтологий и новых способов классификации [Bowker 2000]. Использование Wikidata как ресурса, который не связан с языком описания и использует URI для идентификации нужного значения поля метаданных, сталкивается с рядом трудностей, которые связаны с точностью и качеством данных в Wikidata [Zhao 2023]. Часто обсуждаемая неравномерная представленность данных в Wikidata тоже связана с социальным контекстом и условиями создания датасета [Fischer et al. 2023, Zhitomirsky-Geffet and Minster 2023]

Форматы данных

Следующая возможная причина предпочтения, которое исследователи отдают личным или проектным коллекциям, — отсутствие нужных форматов данных на сайтах библиотек, музеев и архивов. Действительно, по разным причинам, в том числе из-за ограничений, связанных с авторским правом, а иногда от того, что тексты в рамках известных канонов лучше известны ученым и лучше документированы, исследователи предпочитают работать с текстами, созданными до второй четверти двадцатого века, то есть с теми текстами, для которых не существует копий, которые легко могут быть обработаны машинным способом. Исключение составляют англоязычные тексты, которые хранятся в собрании Проекта Гутенберг (Project Gutenberg) и представлены, в том числе, в формате Plain Text и тексты, доступ к которым предоставляют некоторые другие проекты и институции. Одним из способов решения этой проблемы является подготовка цифровых академических изданий исследуемых текстов в соответствии со стандартом Text Encoding Initiative[7] с последующей конвертацией в нужные форматы данных, но такое решение является трудозатратным и дорогостоящим. Еще один способ подготовки текста в нужном формате — работа с одной из платформ, таких как Transcribus[8], которые дают возможность автоматического распознавания, транскрибирования и

конвертации текста в нужный формат, как для рукописных, так и для печатных текстов [Nockels et al. 2022]. С другой стороны ряд крупных библиотек и музеев с активно работающими исследовательскими лабораториями публикуют некоторую часть данных в форматах пригодных для машинной обработки [Candela 2023]. Однако даже и в этом случае публикуются скорее датасеты, в которых представлены метаданные цифровых коллекций. Публикация текстов в форматах, предназначенных для машинной обработки, является скорее исключением, чем правилом.

Неравномерная представленность данных разного типа в цифровой коллекции

Как известно, любой корпус не свободен от скоса и нерепрезентативного представления данных [Chomsky 1957, Clear 1992, Raineri and Debras 2020]. Во многих случаях это может привести к неправильному представлению о распределении в популяции. Это явление много раз обсуждалось при анализе естественно-научных коллекций данных [Daru et al. 2028, Meineke and Daru 2021, Daru and Rodriguez 2023] и агрегаторов, публикующих коллекции культурно-значимых и естественно-научных данных [Kizhner et al. 2021, Raja et al. 2022].

Неравномерное распределение культурно-значимых данных появляется в результате анализа результатов скошенной выборки, которая возникает при неравномерной представленности культур разного типа и разных периодов в датасетах или коллекциях. Такую неравномерность несложно оценить, сравнив метаданные для разных географических локаций и для разных периодов в отдельной цифровой коллекции или агрегаторе. Неравномерность представления данных и скос в сторону западных культур и столичных регионов объясняется разными причинами. Чаще всего коллекции культурно-значимых данных являются зеркалом, которое отражает разные периоды в создании инфраструктур, производящих знание (“infrastructures of knowledge-making” [Mak 2014]). Онтологические пропуски возникают при отсутствии знания об объекте (пропуски или нечеткие формулировки в значениях полей метаданных), в то время как эпистемические скосы – результат заполнения полей или переноса текстов из физических каталогов. Предпочтения, которые отдаются формулировкам и номенклатурным терминам в этом случае вызваны политическими и социальными обстоятельствами и академическими традициями. Терминологические предпочтения

или выбор точки зрения при индексации в большой степени определяются социальным влиянием [Bar-Ilan et al. 2010]. Это приводит к тому, что скос в представлении данных появляется в результате субъективных представлений куратора, хранителя или создателя коллекций, вызванных социальными конвенциями эпохи [Ortolja-Baird and Nyhan 2022, Kizhner et al. 2022]. Цифровые коллекции и агрегаторы, которые не сопровождаются указанием на обстоятельства создания данных, вызывают неожиданные эффекты, которые не предполагались создателями коллекций. Такие эффекты возникают в ситуации, когда объект определенного типа оказывается в неожиданных контекстах и приобретает дополнительные значения из-за связи с другими объектами. В случае, когда это делается с помощью алгоритмов доминирующие связи создают основное значение, которое может быть несвойственно этому объекту в других контекстах. Так например, поиск по запросу "Казахстан" в агрегаторе Google Arts and Culture весной 2023 года приводил к получению фотографий, на которых была представлена деятельность космодрома "Байконур" в Казахстане (6 004 из 7,899 изображений или 75% датасета).

Таким образом, социальные предпосылки создания знания во многом определяют ограничения цифровых коллекций как источников данных. В этом контексте очень важно понимать, "кто финансирует проекты оцифровки, какие критерии отбора применяются, как документируются предметы и тексты, какие классификационные схемы используются" [Zaagsma 2022]. Мы не можем обеспечить равномерное распределение данных в цифровой коллекции, но мы можем отнестись к цифровым коллекциям как к первичным источникам [Liu, 2017]. Эпистемическая и научная ценность коллекций как данных — следствие последовательной, долгой и высококвалифицированной работы, основанной на анализе предшествующих практик [Leonelli 2016, Wyatt 2022], в том числе текстологическом и библиографическом анализе, исследованиях в области истории книги, сравнительных медийных исследованиях и критическом анализе цифровой инфраструктуры [Bode 2020b]. Таким образом, возникает необходимость явным образом представить обстоятельства, при которых формировались физические и цифровые коллекции и объяснить "характер и происхождение онтологических пропусков и эпистемических скосов в данных" [Bode 2020], *перевод мой*). Первым шагом на этом пути может быть явное сообщение о контексте, истории и ограничениях создания цифровой коллекции, которое сопровождает публикацию цифровой

коллекции [Houswedell et al. 2020, Bode 2020a, Zaagsma 2022]. Более подробным образом обстоятельства создания и документирования коллекции могут быть зафиксированы в параданных. Процесс создания параданных — важного типа информации об условиях и контексте создания данных — влечет за собой проблемы, похожие на те, с которыми сталкиваются при разработке метаданных. Разнообразие практик и типов данных в гуманитарных исследованиях приводит к сложностям разработки стандартов и принятия решений относительно того, в какой степени и с какой точностью должны быть зафиксированы такие данные. Как и многие другие данные о данных, параданные никогда не будут зафиксированы полностью [Huvila 2022].

Скосы в цифровых коллекциях и источники неравномерного распределения данных

Основные скосы в цифровых коллекциях связаны с преобладанием культур стран с высоким и средним уровнем дохода, то есть географическим смещением в результатах анализа. Следующий скос связан с преобладанием данных, созданных в девятнадцатом и двадцатом веке, то есть со смещением во времени (см., например, [Zhitomirsky-Geffet and Minster 2023]). Скос, связанный с предпочтениями коллекционеров, руководителей экспедиций, кураторов и хранителей является следующим заметным фактором, влияющим на результаты анализа. Это приводит к тому, что предметы “неслучайным образом распределены в пространстве и времени, представляя смещенные (biased) наборы данных истинного распределения’ [Meineke and Daru 2021], перевод мой).

Можно выделить несколько уровней создания источников неравномерности представления данных:

- Физические коллекции (политические, социальные, экономические ограничения, предпочтения коллекционеров и руководителей экспедиций)
- Метаданные физических коллекций (степень изученности коллекций, степень неопределенности при интерпретации музейных предметов)
- Данные о выставках, опубликованные в физических каталогах (политические ограничения, повестка научных исследований, предпочтения куратора)

- Текстовые описания в физических каталогах (повестка научных исследований, академические традиции)
- Оцифрованные предметы (политические, экономические и технические ограничения)
- Опубликованные цифровые коллекции, метаданные цифровых коллекций (субъективность куратора, экономические ограничения, социальные ограничения, такие как наличие квалифицированных сотрудников)

При этом организация данных в датасеты, вследствие процедур, созданных алгоритмическим образом на основе общности некоторых полей метаданных или значений полей метаданных имеет большое значение, учитывая все большее количество датасетов, созданных с участием алгоритмов. Алгоритмы представления данных разных типов и выравнивания исходного распределения могут играть важную роль в искажении истинного распределения, создавая следующий уровень смещения. Например, данные из локаций с малой представленностью оказываются в результатах поиска в Google Arts and Culture и создают впечатление разнообразия культур в датасете, несмотря на малое количество объектов из этих локаций.

Способы оценки неравномерного распределения

Оценить неравномерное распределение связанное с разными форматами представления данных, вариантами записи значений метаданных и разной степенью глубины и точности отображения для каждого типа данных является сложной задачей. Такие исследования только ищут подходы к решению проблемы нормализации метаданных и обнаружения соответствий между данными разного типа. Решение таких задач осложняется тем, что классификационные схемы принятые при поиске соответствия между данными тоже не предлагают единственное возможное решение. Даже в рамках узкой исследовательской области может существовать несколько возможных классификационных схем (например, при датировке керамики в археологических исследованиях). В ситуации, когда надо принять решение об обнаружении соответствия между разными типами объектов из разных культур, эта задача становится еще сложнее.

Особенно трудно оценить уровень искажения в распределении объектов цифровой коллекции по сравнению с исходным уровнем

распределения в популяции артефактов. Поскольку истинный уровень распределения неизвестен, остается полагаться на косвенные способы определения истинного уровня, подобные тем, которые используются, например, в археологии (см., например, [Drennan et al. 2015, Palmisano et al., 2017, Palmisano et al. 2021]). Для обеспечения валидности результатов могут использовать сочетание нескольких косвенных способов определения истинного уровня [Palmisano et al. 2017]. Сложности и проблемы использования косвенных методов могут быть связаны с нечетким/переменчивым соотношением между косвенным показателем и показателями исходного распределения, например, в наблюдениях за динамикой популяции [Bevan and Crema 2021]. Сами данные, собранные для анализа косвенных показателей могут демонстрировать неравномерное распределение и, таким образом опять приведут к скосу в распределении результатов.

Еще один вариант определить истинное распределение для идентификации скоса в цифровой коллекции — сравнить данные цифровой коллекции с агрегированным мнением экспертов для определения распределения, например, по периодам или для пространственного распределения [Daru and Rodriguez 2023].

При отсутствии данных об исходном распределении можно определить скос, используя данные о распределении в физической коллекции объектов [Kizhner et al. 2019, Kizhner et al. 2021]. Такой подход покажет распределение, связанное с производством артефактов, которые были включены в коллекции физических музеев, архивов и библиотек по разным причинам. Эти институции могли, например, выполнять роль центров, которые фиксировали факт существования небольшого поселения или документировали те виды деятельности, которые в нем происходили [Kelly, 2006]. С другой стороны этот подход не покажет распределение, основанное на восприятии канонических произведений, которое можно оценить, например, используя сравнение распределения в цифровой коллекции с распределением в списках канонических произведений художественных текстов, визуальных произведений и музыкальных произведений. Нечеткость и субъективность подобных списков делают задачу поиска распределения и скоса для такого типа распределения сложной проблемой.

Заключение

В настоящий момент нам доступны датасеты, в которых опубликованы самые частые типы объектов с самыми подробными атрибуциями из самых заметных коллекций. Применение алгоритмов, выделяющих доминирующие типы данных, которые показывают самые частые объекты для пространственного распределения, распределения во времени и других типов распределения (например, по жанрам), приведет нас в мир унифицированных культур, где культурное разнообразие сведено к тем типам культур, которые рекомендованы алгоритмами.

Таким образом, можно утверждать, что объекты в цифровых коллекциях распределены неравномерно и подобраны субъективно. Это связано с отбором предметов в физические коллекции, часто под влиянием предпочтений сотрудников или под влиянием обстоятельств, а также субъективным отбором при оцифровке и публикации. Неравномерное распределение предметов связано с национальными идеологиями и повестками, традициями учреждений, принципами комплектования коллекций в конкретном учреждении, академическими традициями и разнообразием точек зрения. Разные уровни или слои неравномерности (неравномерность при оцифровке, неравномерность при заполнении полей метаданных, неравномерность при выборе точек зрения при индексации объектов) дополнительно осложняют анализ. Онтологические пропуски редко объясняются явным образом, и мы не знаем с какими пропусками нам придется иметь дело, когда собираем данные для исследования. Будущие исследования, по-видимому, будут развиваться в сторону определения популяции объектов (baseline), оценки неравномерности (bias), определения онтологических и эпистемических пропусков в данных, выяснения политических, социальных и экономических причин пропусков, определения стадий и этапов создания неравномерности, предложения решений для смягчения неравномерности (mitigating bias).

Адаптируя вопросы, предложенные относительно естественнонаучных цифровых коллекций [Meineke and Daru 2021], можно предложить следующие направления для дальнейших исследований:

- Как объекты с разнообразной семантикой и многочисленными смыслами могут быть преобразованы в ‘открытые наборы данных’?

- В какой мере объекты и метаданные (учетные записи) отражают реальное знание истории культуры и изменений в культуре во времени и пространстве?

- Действительно ли политические, экономические и социальные обстоятельства вводят неравномерность (bias), которая свойственна конкретным коллекциям, регионам, периодам времени и типам объектов?

- Если на формирование коллекций влияют политические обстоятельства и финансирование, то представляют ли коллекции те каноны, которые приняты политиками и финансирующими организациями? Если да, то как слои канонов представлены в наборах оцифрованных культурно-значимых данных?

- На какие периоды времени, регионы и типы объектов должны быть направлены усилия по формированию коллекций и оцифровке коллекций, чтобы смягчить неравномерность?

Даже если использование цифровых коллекций как источников данных не приведет к возможности делать широкие обобщения в ближайшее время, дальнейшие улучшения, приведут к необходимости интерпретировать результаты анализа с учетом обстоятельств создания коллекций. Возможно, в этом случае мы сможем получать более надежные результаты, а не только воспроизводимые результаты.

Примечания

[1] См., например, http://museum-api.pbworks.com/w/page/21933420/Museum%C2%A0APIs
[2] https://www.gutenberg.org/
[3] http://www.perseus.tufts.edu/hopper/
[4] https://www.distant-reading.net/eltec/
[5] https://dh2022.adho.org/
[6] см., например, https://data.rijksmuseum.nl/controlled-vocabularies/download/
[7] https://teibyexample.org/exist/
[8] https://readcoop.eu/transkribus/

Литература

Исследования

Глазунов, Орехов 2020 — Глазунов, Е. В., Орехов Б. В. (2020). Унификация данных Музейного Госкаталога РФ. Сибирский антропологический журнал, 4 (3), 154-168.

Костенко & Козлова 2021 — Костенко, В. В., & Козлова, А. С. (2021). Госкаталог музейного фонда России: первый подход к прикладному анализу данных. Скиф. Вопросы студенческой науки, (9 (61)), 34-38.

Маслинский 2022 — Маслинский К. "О культуре работы с данными в DH, или роль Репозитория открытых данных", Семинар DHRI, DHRI, Сибирский федеральный университет, Красноярск, 16.02.2022. https://www.youtube.com/watch?v=18BUQBh2P5E

Bar-Ilan et al. 2010 — Bar-Ilan, J., Zhitomirsky-Geffet, M., Miller, Y., & Shoham, S. (2010). The effects of background information and social interaction on image tagging. Journal of the American Society for Information Science and Technology, 61 (5), 940-951.

Besser 2002 — Besser, Howard. "Moving from isolated digital collections to interoperable digital libraries.-First Monday 7.6 (2002).

Bevan & Crema 2021 — Bevan, A., & Crema, E. R. (2021). Modifiable reporting unit problems and time series of long-term human activity. Philosophical Transactions of the Royal Society B, 376 (1816), 20190726.

Binding et al. 2023 — Binding, Ceri, and Douglas Tudhope. "Automatic Normalization of Temporal Expressions.-Journal of Computer Applications in Archaeology 6.1 (2023).

Bode 2020a — Bode, K. (2020a). Why you can't model away bias, Modern Language Quarterly, 81: 1.

Bode 2020b — Bode, K. (2020b). The Archive, in The Cambridge Companion to Literature in the Digital Age (Ed. Adam Hammond). Cambridge: Cambridge University Press.

Borgman 2012 — Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078.

Castellano et al. 2021 — Castellano, Giovanna, and Gennaro Vessio. "Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview.-Neural Computing and Applications 33.19 (2021): 12263-12282.

Candela 2023 — Candela, Gustavo. "Towards a semantic approach in GLAM Labs: The case of the Data Foundry at the National Library of Scotland.-Journal of Information Science (2023).

Candela et al. 2023 — Candela, Gustavo, et al. "A Checklist to Publish Collections as Data in GLAM Institutions.-arXiv preprint arXiv:2304.02603 (2023).

Chomsky 1957 — Chomsky, N. (1957). Syntactic Structures. The Hague: Mouton.

Clear 1992 — Clear, J. (1992). Corpus sampling. In Leitner, G. (ed.) New Directions in English Language Corpora. Berlin: Mouton-de-Gruyter, pp. 21–31.

Conde et al. 2021 — Conde, Marcos V., and Kerem Turgutlu. "CLIP-Art: Contrastive pre-training for fine-grained art classification."Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

da Silva 2022 — da Silva, Camila. "The Ongoing Translation of the Getty Art & Architecture Thesaurus® into Portuguese: An Art Information Access and Retrieval Tool for Cultural Institutions in Portuguese-Speaking Countries.-Getty Research Journal 16.1 (2022): 209-225.

Daru et al. 2018 — Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J., ... & Davis, C. C. (2018). Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytologist, 217 (2), 939-955.

Daru & Rodriguez 2023 — Daru, B. H., & Rodriguez, J. (2023). Mass production of unvouchered records fails to represent global biodiversity patterns. Nature Ecology & Evolution, 1-16.

Doerr et al. 2014 — Doerr, Martin, et al. "Realizing lessons of the last 20 years: A manifesto for data provisioning & aggregation services for the digital humanities (a position paper).-D-lib magazine 20.7/8 (2014).

Drennan et al. 2015 — Drennan, Robert D., Berrey, C. Adam, and Christian E. Peterson. Regional settlement demography in archaeology. ISD LLC, 2015.

Fischer et al. 2023 — Fischer, F., Blakesley, J., Wojcik, P., & Jäschke, R. (2023). Preface: World Literature in an Expanding Digital Space. Journal of Cultural Analytics, 8 (2).

Gnoli 2011 — Gnoli, Claudio. "Animals belonging to the emperor: enabling viewpoint warrant in classification.-Subject access: Preparing for the future 42 (2011): 91.

Harpring 2013 — Harpring, Patricia. Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works. Getty Publications, 2013.

Hauswedell et al. — Hauswedell, T., Nyhan J., Beals M., Terras, M., Bell E. (2020). Of global reach yet of situated con- texts: an exam- ination of the implicit and explicit selection criteria that shape digital archives of historical newspapers. Archival Science, 20: 139–65.

Havens 2022 — Havens, Lucy, et al. "The Library Catalogue as Dataset: Exploring Data Science Approaches to Analyse Collections at Scale."(2022).

Huvila 2022 — Huvila, Isto. "Improving the usefulness of research data with better paradata.-Open Information Science 6.1 (2022): 28-48.

Kelly 2006 — Kelly, L. (2006). Measuring the impact of museums on their communities: The role of the 21st century museum. Intercom, 2 (4).

Kizhner et al. 2021 — Kizhner, I., Terras, M., Rumyantsev, M., Khokhlova, V., Demeshkova, E., Rudov, I., & Afanasieva, J. (2021). Digital cultural colonialism: measuring bias in aggregated digitized content held in Google Arts and Culture. Digital Scholarship in the Humanities, 36(3), 607-640.

Kizhner et al. 2022 — Kizhner, Inna, et al. "The Culture of the Very Rich and Very Poor: Do Digital Museum Collections Tell us Anything about Jewish Culture.-Studies in Digital History and Hermeneutics 43 (2022).

Liu 2017 — Liu, A. Toward Critical infrastructure Studies: Digital Humanities, New Media Studies, and the Culture of Infrastructure / A. Liu. — University of Connecticut, 23 February, 2017

Mak 2014 — Mak, B. (2014). Archaeology of a digitization. Journal of the Association for Information Science and Technology, 65(8): 1515–26.

Meineke & Daru 2021 — Meineke, E. K., & Daru, B. H. (2021). Bias assessments to expand research harnessing biological collections. Trends in Ecology & Evolution, 36 (12), 1071-1082.

Nockels et al. 2022 — Nockels, J., Gooding, P., Ames, S., Terras, M. (2022). Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. Archival Science, 22 (3), 367-392.

Ortolja-Baird & Julianne 2022 — Ortolja-Baird, Alexandra, and Julianne Nyhan. "Encoding the haunting of an object catalogue: on the potential of digital technologies to perpetuate or subvert the silence and bias of the early-modern archive.-Digital Scholarship in the Humanities 37.3 (2022): 844-867.

Palmisano et al. 2017 — Palmisano, A., Bevan, A., & Shennan, S. (2017). Comparing archaeological proxies for long-term population patterns: An example from central Italy. Journal of Archaeological Science, 87, 59-72.

Palmisano et al. 2021 — Palmisano, A., Bevan, A., Kabelindde, A. et al. Long-Term Demographic Trends in Prehistoric Italy: Climate Impacts and Regionalised Socio-Ecological Trajectories. J World Prehist **34**, 381–432 (2021).

Raineri & Debras 2020 — Raineri, S. and Debras, C. (2020). Corpora and representativeness: where to go from now? CogniTextes. 2019, n.p. https: //journals.openedition.org/cognitextes/1671

Raja et al. 2022 — Raja, N. B., Dunne, E. M., Matiwane, A., Khan, T. M., Nätscher, P. S., Ghilardi, A. M., & Chattopadhyay, D. (2022). Colonial history

and global economics distort our understanding of deep-time biodiversity. Nature ecology & evolution, 6 (2), 145-154.

Rei 2023 — Rei, Luis, et al. "Multimodal metadata assignment for cultural heritage artifacts.-Multimedia Systems 29.2 (2023): 847-869.

Sköld et al. 2023 — Sköld, O., Kaiser, J., Andersson, L., Huvila, I., & Liu, Y. H. (2023). Facilitating Data Re-use by Better Understanding Paradata, DHNB2023, Book of Abstracts, pp. 48-51.

Star & Griesemer 1989 — Star, Susan Leigh, and James R. Griesemer. "Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39.-Social studies of science 19.3 (1989): 387-420.

Tolfo et al. 2021 — Tolfo, Giorgia, et al. "Hunting for Treasure: Living with Machines and the British Library Newspaper Collection.".-Studies in Digital History and Hermeneutics (2021): 25.

Vezina et al. 2022 — Vezina, Brigitte, et al. "Towards Better Sharing of Cultural Heritage—An Agenda for Copyright Reform: A Creative Commons Policy Paper."(2022).

Wallace & Euler 2020 — Wallace, A., Euler, E. Revisiting Access to Cultural Heritage in the Public Domain: EU and International Developments. IIC **51**, 823–855 (2020).

Wallace & McCarthy — Wallace, A., McCarthy, D. Survey of GLAM open access policy and practice, 2018-2023, https://docs.google.com/spreadsheets/d/1WPS-KJptUJ-o8SXtg00llcxq0IKJu8eO6Ege_GrLaNc/edit#gid=1409426267

Wyatt 2022 — Wyatt, S. (2022). Critical (big) data studies. In The Necessity of Critique: Andrew Feenberg and the Philosophy of Technology (pp. 127-142). Cham: Springer International Publishing.

Zhao 2023 — Zhao, Fudie. "A systematic review of Wikidata in Digital Humanities projects.-Digital Scholarship in the Humanities 38.2 (2023): 852-874.

Zhitomirsky-Geffet 2019 — Zhitomirsky-Geffet, Maayan. "Towards a diversified knowledge organization system: An open network of inter-linked subsystems with multiple validity scopes.-Journal of Documentation 75.5 (2019): 1124-1138.

Zhitomirsky-Geffet 2023 — Zhitomirsky-Geffet, Maayan, and Sara Minster. "Cultural information bubbles: A new approach for automatic ethical evaluation of digital artwork collections based on Wikidata.-Digital Scholarship in the Humanities 38.2 (2023): 891-911.