ПРЕОБРАЗОВАНИЕ БИБЛИОТЕКИ В ЦИФРОВУЮ ИССЛЕДОВАТЕЛЬСКУЮ ИНФРАСТРУКТУРУ: СОЗДАНИЕ KBLAB В НАЦИОНАЛЬНОЙ БИБЛИОТЕКЕ ШВЕЦИИ

TRANSFIGURING THE LIBRARY AS DIGITAL RESEARCH INFRASTRUCTURE MAKING KBLAB AT THE NATIONAL LIBRARY OF SWEDEN*

Лав Бёрьесон (Love Börjeson), *KBLab, Национальная библиотека Швеции*
Крис Хаффенден (Chris Haffenden), *Упсальский университет, дисциплинарная область гуманитарных и социальных наук, факультет искусств, кафедра истории науки и идей; KBLab, Национальная библиотека Швеции*
Мартин Мальмстен (Martin Malmsten), *KBLab, Национальная библиотека Швеции*
Фредрик Клингволл (Fredrik Klingwall), *KBLab, Национальная библиотека Швеции*
Эмма Ренде (Emma Rende), *KBLab, Национальная библиотека Швеции*
Робин Курц (Robin Kurtz), *KBLab, Национальная библиотека Швеции*
Фатон Рекатати (Faton Rekathati), *KBLab, Национальная библиотека Швеции*
Хиллеви Хэгглёф (Hillevi Hägglöf), *KBLab, Национальная библиотека Швеции*
Юстина Сикора (Justyna Sikora), *KBLab, Национальная библиотека Швеции*

Реферат. В статье рассказывается о создании лаборатории данных KBLab в Национальной библиотеке Швеции (KB). В связи с растущим спросом на предоставление широкого доступа к коллекциям библиотек и других учреждений, занимающихся хранением предметов культурного наследия, организация деятельности в форме лаборатории и применение институционального опыта в области наук о данных являются уместными мерами реагирования. В первой части статьи предлагается оценочное обсуждение работы, связанной с созданием KBLab и как физического, и как цифрового пространства, позволяющего исследователям использовать цифровые фонды KB в немыслимых ранее масштабах. Авторы объясняют, как функционирование лаборатории согласуется с более широкой миссией KB как национальной библиотеки, а также подробно останавливаются на дизайне технической структуры и процессах координации исследований, необходимых для работы лаборатории. Во второй части рассматривается использование в рамках KBLab коллекций библиотеки как источника данных для создания высококачественных моделей искусственного интеллекта для шведского языка, что представляет собой востребованный аспект современной цифровой исследовательской инфраструктуры. Авторы изучают данную опытно-конструкторскую работу в контексте неравномерного охвата технологиями искусственного интеллекта небольших языков и обсуждают, как модели лаборатории способствовали формированию подобной инфраструктуры для шведского языка. В заключение поднимается вопрос о возможностях и проблемах, связанных с продолжением разработки искусственного интеллекта на базе библиотеки, что было инициировано в KBLab.

Ключевые слова: библиотечные лаборатории, цифровая исследовательская инфраструктура, национальные библиотеки, коллекции как базы данных, разработка ИИ.

Введение

В эпоху больших данных к библиотекам предъявляют новые требования [1]. Мир становится все более восприимчивым к процессам датафикации, все больше аспектов повседневной жизни,

* Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden / L. Börjeson, C. Haffenden, M. Malmsten, F. Klingwall, E. Rende, R. Kurtz, F. Rekathati, H. Hägglöf, J. Sikora // SocArXiv Papers. 2023. March 25. DOI: 10.31235/osf.io/w48rf.

которые ранее не поддавались количественной оценке, преобразуются в данные [2], поэтому библиотека как учреждение культурного наследия вынуждена вступить в период творческой трансформации. Отчасти речь идет о разработке методов сбора огромного количества материалов, создаваемых в Интернете, и изучении устойчивых способов описания и сохранения этих веб-архивов для будущих пользователей [3]. Однако это также предполагает наличие стратегий для удовлетворения нужд пользователей в настоящем, особенно новаторских потребностей цифровой науки [4]. Исследователи в области гуманитарных и социальных наук, использующие цифровые подходы, сегодня рассчитывают, что смогут проводить анализ библиотечных коллекций в немыслимых ранее масштабах (см., например, [5]). Подобные ожидания, наиболее очевидные в случае научно-исследовательских и национальных библиотек, работающих с обязательным экземпляром, создают особые проблемы для информационных систем, которые исторически отдавали предпочтение аналоговым объектам и использованию отдельных копий. Как библиотеки реализуют предоставление доступа к своим фондам в русле концепции «коллекции как данные», когда так много из того, что лежит в основе их социотехнических наработок, базируется на предоставлении физических экземпляров [6; 7]?

В настоящей статье указанная проблема рассматривается через призму лаборатории данных (Data lab) как организационной формы. В последнее десятилетие, столкнувшись с растущими потребностями в предоставлении доступа к фондам с возможностью машинной обработки документов, университетские и национальные библиотеки отреагировали посредством создания подобных лабораторий. Типичными примерами являются: лаборатория LC Labs в Библиотеке Конгресса США, лаборатория British Library Labs в Британской библиотеке и лаборатория цифровых гуманитарных наук Yale Digital Humanities Lab в Йельском университете. В широком смысле такие проекты предполагают создание внутренней платформы, в рамках которой профессиональный опыт специалистов по данным может быть использован для решения проблем оцифровки и содействия новым формам цифровых исследований. В представленной работе обсуждается развертывание такой лаборатории в стенах библиотеки на примере KBLab в Национальной библиотеке Швеции (KB). В первой части статьи подробно описана организация инфраструктуры, необходимой для того, чтобы сделать цифровые коллекции KB доступными для крупномасштабного анализа, а также созданные в KBLab практические и технические условия. Во второй части объясняется, как использование коллекций в качестве основы для разработок в области применения искусственного интеллекта (ИИ) оказалось основополагающим для преобразования библиотеки в цифровую исследовательскую инфраструктуру. Несмотря на то что конкретные характеристики KBLab специфичны для шведского контекста, авторы приводят более общие аргументы, актуальные для широкой международной аудитории библиотечных специалистов, исследователей цифровых технологий и представителей власти. Таким образом, статья представляет собой ответ на недавно вышедший манифест, поощряющий рост новых лабораторий в секторе культурного наследия, а именно «Откройте GLAM-лабораторию» [8], а также разъясняет ценность разработки искусственного интеллекта, ориентированного на библиотеки, как общественного блага.

KBLab как точка доступа к коллекциям для исследователей

Во вводном разделе изложены практические и организационные предпосылки, повлиявшие на создание лаборатории данных в Национальной библиотеке Швеции. Каким образом проект KBLab согласуется с более широкой миссией KB как национальной библиотеки и интегрирован в ее реализацию? За счет чего фонды национальной библиотеки особенно хорошо подходят для той работы, которая возможна в подобной лаборатории? И что необходимо для создания доступа к материалам в лабораторной среде? Отвечая на эти вопросы, авторы разъясняют детали, которые требуются для понимания особенностей разработки ИИ в библиотеке.

Библиотечные коллекции как данные

Организация широкого доступа к коллекциям занимает центральное место в функциональном назначении KB как бюджетного учреждения. Обязательства библиотеки перед исследовательским сообществом в этом отношении отражены в законодательном акте, определяющем ее основную миссию, где в первом же абзаце KB определяется не только как национальная библиотека, но и как национальная исследовательская инфраструктура [9, p. 1421]. Хотя конкретные задачи, относящиеся к данной миссии, т. е. собирать, описывать, сохранять и делать доступными публикации, связаны в первую очередь, с общим содействием демократическому развитию, закон непосредственно оговаривает, что указанные задачи служат цели поддержки научной деятельности в Швеции. В этом смысле на KB лежит юридически закрепленное обязательство удовлетворять изменяющиеся потребности исследователей. На практике, учитывая цифровизацию медиасреды, это означает внедрение цифровых технологий в национальную исследовательскую инфраструктуру и затем

превращение ее в цифровую исследовательскую инфраструктуру. Как будет показано в статье, KBLab включает в себя ключевые элементы обоих направлений деятельности.

Фонд KB характеризуется большим объемом. Закон Швеции об обязательном экземпляре был принят в 1661 г. как метод официальной цензуры, поскольку вынуждал издателей предоставлять копию каждого произведения государству на утверждение до опубликования, однако он также способствовал становлению национальной библиотеки как гаранта сохранности культурного наследия для будущих поколений. Согласно требованиям закона, копия каждой публикации, изданной на шведском языке, должна быть передана в библиотеку; с 1979 г. под это требование подпадают аудиовизуальные материалы и печатные издания, а с 2015 г. — как минимум часть электронных публикаций [10]. Также коллекции содержат культурную продукцию, созданную в условиях разнообразного медиапространства: от газет, журналов, книг, научной периодики и правительственных отчетов до радиопередач, телевизионных шоу и компьютерных игр. Чтобы дать представление о масштабах фонда, отметим, что одни только физические архивные коллекции KB в настоящее время насчитывают более 18 млн единиц хранения.

Несмотря на то что небольшую часть фонда еще предстоит оцифровать, уже накоплен большой объем цифровых материалов, что делает подход «коллекции как данные», изложенный Томасом Падиллой (Thomas Padilla) и другими авторами [11; 12], весьма актуальным. Внедрение этой концепции открывает большие возможности, но также влечет за собой серьезные инфраструктурные проблемы для сектора GLAM. Так, к ее достоинствам относится создание высококачественных баз данных по гуманитарным наукам для конкретных языков, что потенциально позволит исследователям анализировать содержимое цифровых коллекций в невиданных ранее масштабах и часто неизвестными ранее способами. Существование таких баз данных особенно ценно при разработке инструментов ИИ для менее распространенных языков — момент, к которому мы еще вернемся.

Однако предоставление доступа к данным по гуманитарным наукам — это не самая простая задача. Коллекции библиотеки имеют свою особую историю, сформировавшую их архивную форму, при этом трансформация предметов в данные предполагает дальнейшее управление этими данными. В качестве примера рассмотрим использование программного обеспечения оптического распознавания символов (OCR) в процессе создания цифровых копий материалов [13]. Согласно положениям шведского закона об обязательном экземпляре предпочтение ранее отдавали физическим копиям, а не цифровым, поэтому бумажные газеты отправляли в KB, где их затем оцифровывали. Помимо некоторых ошибок OCR, связанных с неверным распознаванием шведских слов, последствием оцифровки является потеря различных аспектов метаданных, которые мы часто воспринимаем как должное. Так, в результате этого процесса у нас может появиться цифровая копия, однако мы получим мешанину из текстовых блоков, ведь неизвестно, какие блоки связаны между собой и формируют часть одной и той же статьи, какие статьи составляют часть одного и того же раздела, какие тексты являются редакционным контентом, а какие — рекламой. Разумеется, можно использовать технологии машинного обучения, чтобы попытаться реконструировать газету, но это сложная и трудоемкая задача [14]. Создание коллекций, поддающихся компьютерному анализу, является специализированной задачей, которая часто требует значительных усилий с точки зрения очистки и обработки данных, т. е. гуманитарные данные не поступают в библиотеки уже в готовом виде (см. [15]).

Именно в этом конкретном контексте появилась KBLab. С одной стороны, растут запросы ученых в области гуманитарных и социальных наук, которые используют цифровые подходы и хотят иметь доступ к цифровым коллекциям для проведения крупномасштабного анализа. Как показало пилотное исследование, которое заложило основу для открытия лаборатории, ученые, агентства, предоставляющие финансирование, правительственные гранты на исследования также все чаще толкают науку в направлении интенсивного использования данных, чтобы продвигать цифровую науку [16, p. 29]. С другой стороны, существует техническая сложность, связанная с созданием целой инфраструктуры, которая способна фактически с нуля обеспечить доступ к этим коллекциям как к данным. Подготовка высококачественных массивов данных, пригодных для исследований, может оказаться трудоемкой задачей. Далее рассматриваются конкретные шаги, направленные на решение этой проблемы путем создания лаборатории в KB.

Проектирование технической инфраструктуры

Когда библиотека официально инициировала проект по созданию лаборатории данных в 2019 г., были определены две конкретные группы пользователей, каждая со своими особыми целями. В самой библиотеке KBLab задумывалась как внутренний ресурс для разработки методов и инноваций в области ИИ, т. е. средство предоставления персоналу и руководству актуальных знаний о потенциале автоматизации различных рабочих процессов библиотеки. Лаборатория должна была найти свою нишу среди существующих учреждений и информационной среды, чтобы стать обще-

Рис. 1. Визуализация лабораторной среды для открытых исследований, графический интерфейс KBLab

признанной инфраструктурой, поддерживающей проведение цифровых исследований. В краткосрочной перспективе это должно было удовлетворить потребности двух крупных проектов в области цифровых гуманитарных и социальных наук, финансируемых Шведским исследовательским советом: «Аналитика государства всеобщего благосостояния. Анализ текста и моделирование шведской политики, СМИ и культуры, 1945—1989» (на базе Университета Умео)[1] и «Генерирование смысла: динамика публичного дискурса по проблеме миграции» (на базе Университета Линчёпинга)[2]. То, что оба проекта связаны с проведением крупномасштабного анализа материалов середины и конца XX в. из коллекций КВ (в основном газет, но также художественной литературы и периодических изданий), оказало существенное влияние на техническое и организационное развитие лаборатории. Поскольку исследования предполагали анализ материалов, все еще защищенных авторским правом, первоначальная задача заключалась в разработке вычислительной инфраструктуры для обеспечения локального доступа к публикациям на территории библиотеки.

Отправной точкой в решении этой задачи было твердое убеждение в том, что лаборатория данных при библиотеке должна предлагать экспериментальный доступ к коллекциям. Чтобы заслужить звание лаборатории, разработчики стремились предоставить пользователям среду, которая поддерживала бы и поощряла проведение исследований. Разумеется, одним из способов сделать цифровые коллекции доступными для дальнейших исследований является создание предварительно определенных наборов данных, которые затем могут быть выложены в целях анализа или использования другими способами. Однако помимо того, что требования соблюдения авторского права в данном случае препятствовали такому подходу, специалисты КВ стремились создать среду, которая поддерживала бы открытые критические исследования, вместо того чтобы ограничивать доступ к коллекциям, предлагая уже предопределенные наборы данных. Основываясь на собственном опыте исследовательского процесса, не всегда линейного, а в действительности зачастую тангенциального и продиктованного случайными открытиями, которые выходят далеко за рамки первоначальной задачи исследования [17], создатели решили спроектировать KBLab как инфраструктуру, в которой посетителям предоставлена возможность свободно изучать материалы. Как только исследовательский проект зарегистрирован в лаборатории, участникам предоставляется неограниченный доступ к цифровым коллекциям базы знаний, чтобы они могли исследовать и разрабатывать свои собственные наборы данных в результате работы с библиотечными фондами.

Связанная с этим техническая проблема заключалась в том, чтобы обеспечить проведение исследований без угроз безопасности. Решение состояло в том, чтобы организовать непрямой доступ к документам через интерфейс прикладных программ (API). Исследователи имеют возможность проводить поиск в цифровых коллекциях базы знаний через API-лаборатории, получая результаты в виде исходных данных, без лишнего риска раскрытия баз данных библиотеки через прямой доступ к самим файлам [18]. Результаты представлены в виде файлов JSON, что является формой данных, с одной стороны, отражающей долгую историю взаимодействия КВ со связанными дан-

ными, а с другой — особенно уместной для инфраструктуры цифровых исследований, поскольку она является машиночитаемой [19]. Еще одним ключевым аспектом связанной модели данных, лежащей в основе дизайна лаборатории, является наличие унифицированных идентификаторов ресурсов (URI) в лабораторной среде. Наличие стабильных и неизменных URI для архивных материалов позволяет исследователям вернуться к тому или иному месту в коллекциях, а также при необходимости продемонстрировать, что их результаты воспроизводимы. Создав API и информационную модель, которая связывает данные цифрового архива, в KBLab удалось обеспечить программный доступ к коллекциям библиотеки, предоставив пользователям автономию для управления своими исследовательскими процессами.

В дополнение к возможности поиска материалов через API специалисты KB также разработали графический пользовательский интерфейс (GUI) для лабораторной среды (рис. 1). Это способствует повышению функциональности лаборатории как исследовательской инфраструктуры различными, но пересекающимися путями. Во-первых, интерфейс предоставляет средства проверки: при доступе к материалу в визуальной форме исследователи могут проверять свои результаты, а также осуществлять по ним навигацию. Во-вторых, он дает ученым в области гуманитарных и социальных наук, не имеющим навыков программирования, возможность доступа к материалам и пользования ими в лабораторной среде. Данный метод особенно актуален для междисциплинарных проектов, которые предполагают объединение подходов наук о данных с более традиционными навыками внимательного чтения и изучения документов в рамках различных гуманитарных дисциплин. Это также уместно для осуществления смешанного анализа, который, помимо проведения крупномасштабного вычислительного анализа, включает исследование визуальных аспектов материала (следовательно, ученый также должен иметь возможность видеть отдельный объект, а не его текстовое содержание). В-третьих, GUI позволяет аннотировать материал, что может оказаться важным элементом в проектах, использующих машинное обучение путем тренировки моделей на основе коллекций. В интерфейсе есть специальная функция, которая помогает пользователям осуществлять аннотирование в соответствии с выбранными ими метками, а затем извлекать определенный текст, который был ранее аннотирован (рис. 2).

Лабораторный графический интерфейс сделан доступным для исследователей за пределами самой лаборатории через прототип службы под названием бета-лаборатория (betalab, см. https://betalab.kb.se/). С одной стороны, служба является частью процесса регистрации исследовательских проектов, которым был предоставлен доступ к использованию KBLab. Иными словами, прежде чем получить физический доступ к помещению лаборатории, исследователи могут использовать бета-лабораторию для тестирования и адаптации к лабораторной среде, в некоторых случаях они могут даже разработать и подготовить сценарии для за-

Рис. 2. Аннотирование и извлечение данных исследования в лабораторной среде

пуска на месте, как только доступ будет получен (это может значительно сэкономить время для географически разрозненных проектов). С другой стороны, бета-лаборатория также используется как точка доступа к тем частям цифровых коллекций, которые доступны в лаборатории, но не подпадают под ограничения в силу авторского права. Эти открытые данные включают исторические газетные материалы до 1906 г., официальные отчеты правительства Швеции и различные парламентские документы. Если, например, в рамках исследовательского проекта ученым нужно получить доступ к газетным материалам в машиночитаемой структурированной форме для изучения криминальных репортажей XIX в., доступ к этим данным предоставляется через бета-лабораторию. В этом смысле он является важным дополнением к основной части лабораторной среды, доступ к которой можно получить в помещении KBLab.

Для исследовательских проектов, в рамках которых необходимо использовать KBLab для крупномасштабного компьютерного анализа тех частей коллекций, которые защищены авторским правом, была создана компьютерная лаборатория в помещении одного из зданий KB в Стокгольме. Функционирование этого физического воплощения лаборатории в качестве исследовательской инфраструктуры в значительной степени зависит от вычислительной мощности. Так, согласно положениям шведского законодательства об авторском праве в настоящее время участники проектов не могут перемещать данные за пределы лаборатории для внешней обработки, поэтому библиотека должна располагать достаточным объемом внутренних вычислительных ресурсов для удовлетворения потребностей исследователей. В связи с этим была построена локальная вычислительная инфраструктура с тремя уровнями: оборудованные мощными компьютерами рабочие места в лаборатории; серверная среда для проведения вычислений и осуществления доступа к материалам через API; два сервера NVidia DGX A100 для анализа, требующего больших вычислительных ресурсов. (Позднее библиотеке также был предоставлен доступ к суперкомпьютерной инфраструктуре EC для поддержки собственных разработок, что будет рассмотрено далее.)

Руководящие принципы, которые легли в основу технической разработки этого инженерного решения, — прагматизм, гибкость и стремление обеспечить автономию для исследователей. Так, рабочие места в компьютерной лаборатории были оборудованы системой Ubuntu на базе Linux, поскольку это позволяет пользователям создавать и контролировать свою программную среду в соответствии с собственными потребностями и предпочтениями. Чтобы позволить исследователям управлять резервными копиями своего кода и незавершенной работой, аналогичным образом была создана так называемая гит-лаборатория, внутренняя серверная функция *git*. При этом было решено начать с приобретения пользовательского, а не корпоративного оборудования для лаборатории: отчасти в силу (относительно) ограниченных ресурсов, которые были в распоряжении создателей, но также потому, что это позволило им действовать оперативно и своевременно адаптироваться к меняющимся потребностям исследователей. Работа, связанная с разработкой всего комплекса, базировалась на имеющемся опыте персонала KB в области ИТ-архитектуры и проектирования систем; без привлечения опытного и креативного ИТ-архитектора создание KBLab было бы невозможно.

Координирование исследований

Поскольку лаборатория создавалась как точка доступа к коллекциям, другой важный организационный вопрос, который следовало тщательно обдумать, заключался в том, как координировать исследования. Важным аспектом подготовительной работы стало определение принципов и процедур, определяющих порядок предоставления доступа в компьютерную лабораторию. Спрос на использование KBLab среди исследователей постоянно превышал имеющиеся у библиотеки возможности [20, p. 9], как же следовало распределять ограниченное количество мест на рабочих станциях? Чтобы решить эту проблему справедливым способом, который соответствует ценностям и задачам KB как государственного учреждения, мы сделали подачу заявления в лабораторию частью более широкого процесса библиотеки по управлению приложениями для исследований и разработок[3]. Так, ученые, заинтересованные в сотрудничестве с KBLab, должны сначала представить краткое описание проекта, указав, что они хотели бы сделать в рамках предполагаемого эксперимента. Затем заявка проходит первоначальную проверку для подтверждения, что инициатива действительно включает элементы исследования, т. е. что существуют вопросы и гипотезы, поддающиеся экспериментальному изучению, прежде чем представители библиотеки перейдут к принятию решения о конкретных условиях, в соответствии с которыми возможно сотрудничество. Здесь важно отметить, что библиотека не выносит суждений о содержательном наполнении исследовательского предложения, кроме подтверждения наличия предмета исследования и определения его принципиальной обоснованности.

Проблема стабильного финансирования является центральной для функционирования любой лаборатории [8, p. 129], и это также влияет на то, как организуется деятельность KBLab. Изначально расходы на лабораторию были компенсированы

за счет сочетания внутренних ресурсов библиотеки и упомянутых выше внешних поступлений, однако концепция лаборатории состоит в том, что осуществляемые в ней исследовательские проекты должны быть самофинансируемыми, т. е. исследователи оплачивают накладные расходы для покрытия текущих затрат (технических и административных) при использовании лаборатории в соответствии с общей шведской практикой использования исследовательской инфраструктуры. Согласно существующим схемам финансирования академических изысканий, шведские ученые должны включать смету затрат на использование лаборатории в заявку, которую они подают в спонсорские организации, такие как Шведский исследовательский совет и банк Riksbankens Jubileumsfond. При этом необходимо координировать подачу заявки на получение рабочего места в лаборатории с процессом подачи заявки на финансирование в эти организации.

Преимущество данного подхода состоит в том, что он служит механизмом контроля качества: предоставляя место в лаборатории тем соискателям, чьи идеи получили финансирование в результате конкурентного, коллегиально рецензируемого процесса, можно иметь гарантировнно высокий уровень одобренных проектов. Однако потенциальный недостаток связан с тем, что предпочтение могут отдавать более крупным проектам, предложенным авторитетными учеными, в ущерб небольшим по масштабу инициативам малоизвестных исследователей. В связи с этим библиотека прибегает к прагматичному анализу экономической эффективности при рассмотрении каждого проекта, что может позволить в определенных случаях обойтись без платы за использование лаборатории. Так, если одним из последствий реализации замысла станет значительное улучшение инфраструктуры учреждения, то, возможно, получится найти приемлемое для всех сторон решение. Типичным примером ситуации, когда потенциальные преимущества для библиотеки перевешивают любые накладные расходы, являются проведенные в лаборатории различные магистерские работы в области машинного обучения, в рамках которых было изучено, как модели ИИ могут способствовать повышению доступности библиотечных коллекций [21; 22].

Еще один параметр, влияющий на оценку накладных расходов, — сопоставление уровня компетентности представителей проектной группы в области наук о данных и сложности предлагаемого исследования. Основная проблема здесь заключается в поиске баланса между опытом в области ИИ и машинного обучения с одной стороны и более традиционными качественными компетенциями в области гуманитарных и социальных наук — с другой [23; 24]. Исходя из нашего опыта, привлечение внешних специалистов для крупномасштабного анализа данных является наименее эффективным подходом. Отчасти проблема состоит в том, что подобное решение обесценивает жизненно важный технический труд специалистов, которые не получают должного внимания и признания. Кроме того, он создает ситуацию, когда исследователи в области гуманитарных и социальных наук публикуют работы, в которых они не до конца понимают ни используемые методы, ни полученные результаты.

Учитывая эти соображения, рекомендуется включать в реализуемые на базе KBLab проекты компетентных специалистов в области наук о данных, а также использовать их комментарии и предложения на всех этапах исследовательского процесса. На практике это означает, что библиотека неохотно предоставляет лабораторные помещения командам, которым не хватает необходимых технических навыков, вместо этого перенаправляя их в другие организации Швеции, такие как различные центры цифровых гуманитарных исследований, которые могут оказать более широкую помощь. Когда же речь идет об отвечающих требованиям инициативах, библиотека предлагает схему оплаты накладных расходов, которая корректируется в зависимости от технической сложности и условий того или иного проекта: от стандартной ставки, включающей первоначальную поддержку и консультации по использованию лаборатории, до более высоких уровней, когда от сотрудников лаборатории потребуются серьезные усилия. В каждом случае выяснение конкретных потребностей и условий предполагает постоянный диалог между исследователем и персоналом лаборатории.

Как только проект получил место в лаборатории и гарантии финансирования, библиотека готова принять его в разработку. Данный процесс основан на модели открытого исследования, упомянутой выше при обсуждении технических установок. Вводная ориентация подробно разъясняет условия пользования лабораторией, затем исследователи подписывают пользовательское соглашение, определяющее правовые нормы доступа и использования данных, предоставляемых в KBLab, а также получают копию Правил поведения в библиотеке (см. Приложение 1). Затем следует практическая часть, в ходе которой исследователям демонстрируют, как получить доступ к данным через API, как управлять текущими результатами и какие из предлагаемых инструментов могут оказаться особенно полезными. Пройдя необходимую подготовку, исследователи готовы к автономной деятельности: помимо консультаций с персоналом в случае возникновения проблем, они могут приступить к работе с коллекциями в KBLab в соответствии со своими интересами.

Последний момент, который стоит отметить в связи с процессом координирования исследований, — это то, что квалифицированный персонал здесь играет центральную роль, от него зависит эффективность работы. Указанный процесс предполагает доверительный диалог с пользователями, поэтому важно, чтобы и заведующий лабораторией, и внутренний координатор, и лицо, ответственное за общее руководство научно-исследовательской деятельностью библиотеки, имели ученые степени и практический опыт проведения научных изысканий. Предоставление исследователям профессиональных и высококачественных услуг предполагает наличие квалифицированных и компетентных специалистов среди сотрудников библиотеки.

Модели на основе коллекций как составляющая цифровой инфраструктуры

Рассмотрев создание лаборатории как физического пространства для доступа к фондам, перейдем к обсуждению использования коллекций в качестве основы для разработки новых цифровых инструментов, которые сами по себе составляют важный элемент исследовательской инфраструктуры. В то время как на количество исследователей, которые могут использовать локальную лабораторию, неизбежно налагаются физические ограничения, создание таких инструментов, которые можно использовать за пределами библиотеки, позволяет добиться гораздо большего охвата. Далее в статье речь идет о работе KBLab по созданию и выпуску моделей на основе коллекций: как цифровые коллекции KB способствовали разработке ИИ на базе библиотеки? В каких контекстах и с какой целью используются лабораторные модели ИИ?

Библиотечные коллекции как набор данных для обучения

За последние пять лет в области ИИ и машинного обучения произошли значительные изменения. Например, выпуск языковых моделей на основе преобразователей, таких как BERT, послужил основой для наращивания мощности и производительности в решении многих задач, связанных с обработкой естественного языка [25]. Однако распространение подобных инструментов ИИ произошло в соответствии с существующей глобальной схемой распределения власти и ресурсов: они оказались далеко не в равной степени доступны для всех языков или заинтересованных лиц. В то время как в рамках Google AI были разработаны специальные модели BERT с передовыми технологическими возможностями для таких языков, как английский и китайский, для работы с другими языками приходится довольствоваться менее мощной многоязычной моделью. Однако, когда у крупных технологических компаний не было коммерческого стимула к обучению моделей ИИ, представители научных кругов и другие стороны, как правило, брали на себя инициативу по созданию современных одноязычных моделей (например, [26; 27; 28]). В случае «менее востребованных» языков существенными препятствиями для этого стало отсутствие достаточных вычислительных ресурсов и обучающих данных. Так, в Швеции первая одноязычная модель BERT была создана Государственным агентством по трудоустройству с использованием данных исключительно из шведской версии Википедии. Получившаяся модель все же значительно уступает по эффективности англоязычному продукту BERT (хотя и лучше, чем многоязычная модель Google) [29, p. 35—36].

Тем не менее парадигма, в рамках которой создаются современные модели ИИ, позволяет национальным библиотекам и другим учреждениям культуры вносить особый вклад в исследования, особенно в случае малых языков. Так, текущие разработки в области ИИ проходят в русле концепции неконтролируемого обучения, т. е. масштабные алгоритмы, называемые искусственными нейронными сетями, обучаются, подвергаясь воздействию огромных объемов немаркированных обучающих данных, тогда как ранее использовались меньшие объемы (дорогостоящих) аннотированных данных [30]. Таким образом, для учреждений — хранителей больших объемов высококачественных языковых данных появились новые возможности для участия в этом процессе.

Охват и размер упомянутых выше коллекций KB в этом контексте делает их уникальным и ценным ресурсом для создания передовых инструментов с целью развития шведского ИИ. Действительно, тот факт, что посредством обязательного экземпляра KB имеет доступ к языковым данным, которые максимально приближены к отражению демографических характеристик всего населения, означает, что использование библиотечных фондов в качестве массива данных для обучения ИИ имеет важный демократический аспект. При обращении к более широкому и более репрезентативному диапазону данных, чем тот, который доступен другим субъектам (имеющим доступ в основном лишь к тем данным на шведском языке, которые можно найти в Интернете), KB может создавать более эффективные и высококачественные модели ИИ. Учитывая, что эти данные не могут быть перемещены за пределы библиотеки в силу требований законодательства об авторском праве и Общего регламента ЕС по защите персональных данных, это создает веские основания для обучения моделей ИИ на территории KBLab [29, p. 35].

Создание и распространение ИИ на основе коллекций

В рамках повышения качества шведской инфраструктуры ИИ до мирового уровня цифровые фонды библиотеки использовались для обучения ИИ с момента основания лаборатории в 2019 году. Первый этап разработки инструментов ИИ был сосредоточен непосредственно на работе с текстом с целью улучшения возможностей автоматического анализа текстов на шведском языке в свете последних нововведений в архитектуре трансформеров. Здесь мы обратились к огромным фондам KB, чтобы обучить модель BERT, созданную для шведского языка и способную обрабатывать то, что мы назвали живым языком национального сообщества [31]. С целью обеспечения репрезентативности данных был создан крупный по размеру и разнообразный по составу учебный корпус текстов, в котором использовался большой объем оцифрованных библиотекой архивов газет начиная с 1945 г., а также более свежие онлайн-материалы и тексты из социальных сетей для охвата образцов разговорного языка. Отметим, что адаптация этих текстов для использования в качестве материала для машинного обучения также являлась кропотливым и трудоемким процессом, при этом эффективность обработки данных, в свою очередь, зависела от компетентности сотрудников лаборатории в области программирования. Полученная в результате языковая модель, получившая название KB-BERT, оказалась значительно более эффективной, чем существовавшие модели, и с тех пор она стала стандартной при обработке шведского языка (мы вернемся к вопросу о ее использовании ниже)[4].

В соответствии с мультимодальным направлением текущих инноваций в области ИИ и мультимедийным уклоном последних гуманитарных исследований изыскания KBLab также вышли за рамки исключительно текста. Библиотека воспользовалась преимуществом наличия разнообразных медиаресурсов, хранящихся в архиве: KB располагает обширными коллекциями текстов, изображений, аудио- и видеозаписей на шведском языке в различных форматах, что существенно расширяет диапазон возможностей для обучения новых моделей ИИ. Показательным примером является проект лаборатории по созданию улучшенных инструментов для автоматического распознавания звука (ASR). В ходе этой работы использовались огромные и часто малоисследованные фонды аудиовизуальных материалов XX века. В частности, были задействованы оцифрованные записи национальных и местных радиопрограмм последних двух десятилетий с целью создания корпуса из более чем 1,4 млн часов разговорной речи на шведском языке, включая диалекты всех регионов страны [32]. Собранные материалы послужили в качестве обучающих данных для шведских версий модели wav2vec 2.0, разрабо-

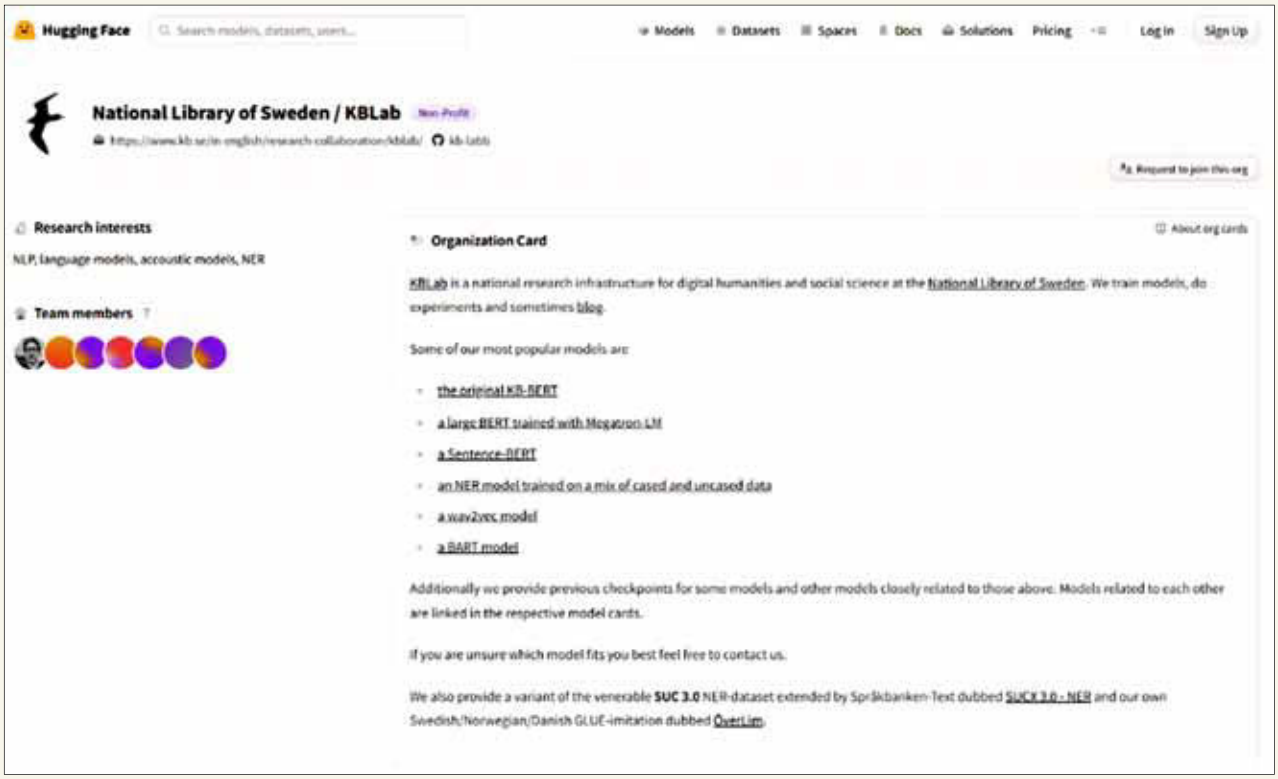

Рис. 3. Шведские модели ИИ, представленные в открытом доступе для скачивания

танной Facebook AI (принадлежит компании Meta Platforms Inc, признанной экстремистской организацией, ее деятельность запрещена на территории Российской Федерации) [33]. Как и в случае с KB-BERT, созданные при этом на основе коллекций модели под названием VoxRex превзошли существующие многоязычные и одноязычные модели в том, что касается преобразования речи в текст[5] [32]. Как будет показано ниже, наличие высокотехнологичных инструментов для преобразования речи в текст, разработанных для шведского языка, создает ряд синергетических эффектов как внутри учреждений культуры, так и за их пределами.

Чтобы инструменты ИИ, которые создаются в KBLab, могли принести пользу как можно большему числу людей, организован открытый доступ к моделям через платформу сообщества исследователей данных HuggingFace[6] (рис. 3). На сегодняшний день пользователи могут загружать 46 моделей и далее экспериментировать с ними в соответствии со своими интересами. В дополнение к указанным моделям KB-BERT и VoxRex отметим канонический шведский SpaCy, классификатор по принципу zero-shot (без обучающих примеров), модели Sentence-BERT (BERT для предложений) и BERT, настроенный для распознавания именованных сущностей (NER), а также шведские версии последних моделей Whisper для ASR, выпущенные OpenAI[7]. Помимо обучения собственных моделей для текстов и аудиозаписей на шведском языке, библиотека также сотрудничала с другими сторонами в разработке мультимодальных инструментов, которые соединяют изображение и текст, чтобы обеспечить новые формы поиска изображений [34]. В рамках прозрачного и ответственного подхода к разработке ИИ сотрудники лаборатории документально фиксируют данные, использованные для обучения моделей, публикуя описания на платформе HuggingFace, сообщения в блогах и научные статьи[8]. Они также делятся кодами через Github[9]. Таким образом, пользователи моделей получают возможность понять, как были созданы эти инструменты, и рассмотреть, как с учетом конкретных ценностей и ключевых положений, содержащихся в данных из коллекций KB, эти модели следует корректировать для использования в тех или иных целях [20, p. 14].

Практическая ценность моделей ИИ, созданных на основе коллекций

Если мы обратимся к изучению пользы от обученных в лаборатории моделей ИИ, то увидим, что они применяются в самых разных контекстах. Первоначальным толчком к их созданию было стремление сделать огромные, но зачастую малоизведанные фонды цифровых материалов библиотеки более доступными. Иными словами, разработка инструментов была призвана помочь библиотеке лучше понимать и описывать собственные коллекции, одновременно расширяя доступ исследователей к документальному наследию. О том, что это удалось сделать, свидетельствуют, например, различные магистерские работы, реализованные в рамках KBLab, показавшие, как KB-BERT можно использовать с целью автоматического обогащения метаданных в цифровом архиве газет (например, [35]). Другим показательным примером является пилотный проект, который ставил целью изучить, как подход к моделированию темы, основанный на шведском аналоге Sentence-BERT, BERTopic, может быть использован для представления автоматических предметных заголовков, которые предлагают более точную навигацию по коллекциям [36]. Однако наиболее интересным явлением, пожалуй, стал цикл позитивной обратной связи, созданный лабораторией при разработке упомянутых выше звуковых данных. Так, коллекции KB позволяют продуцировать новейшие модели ASR; затем эти модели задействуют для преобразования речи в текст, чтобы сделать коллекции пригодными для текстового поиска; а текстовые расшифровки радио- и телевизионных материалов, в свою очередь, можно использовать в качестве новых обучающих данных для следующего поколения более совершенных моделей KBLab. Таким образом, разработки в области ИИ и более качественные метаданные в совокупности применяются для повышения доступности материалов, тем самым делая библиотеку более эффективной исследовательской инфраструктурой.

Помимо библиотеки и исследовательских проектов, базирующихся в лаборатории, модели KBLab оказались полезными для множества научных исследований, а также для организаций, работающих с большими объемами информации за пределами академической среды, как в государственном, так и в частном секторе. Во-первых, KB-BERT теперь используется учеными-медиками для разработки новых методов лечения диабета; в целях автоматической идентификации наличия имплантатов (например, кардиостимуляторов или стентов) у пациентов с пороками сердца перед проведением МРТ-сканирования; для классификации юридических документов [37; 38; 39]. Во-вторых, модели KBLab применяются для автоматизации и оптимизации процессов обработки информации различными государственными органами, включая местные советы, Шведское налоговое агентство (Skatteverket), шведские суды (Domstolsverket), и, совсем недавно, для поддержки государственного управления (сервисный центр Statens)[10]. По мере того как все больше шведских организаций и компаний начинают осознавать возможности ИИ, они все чаще обращаются к библиотечным моделям, предоставляющим легкий доступ к самым совре-

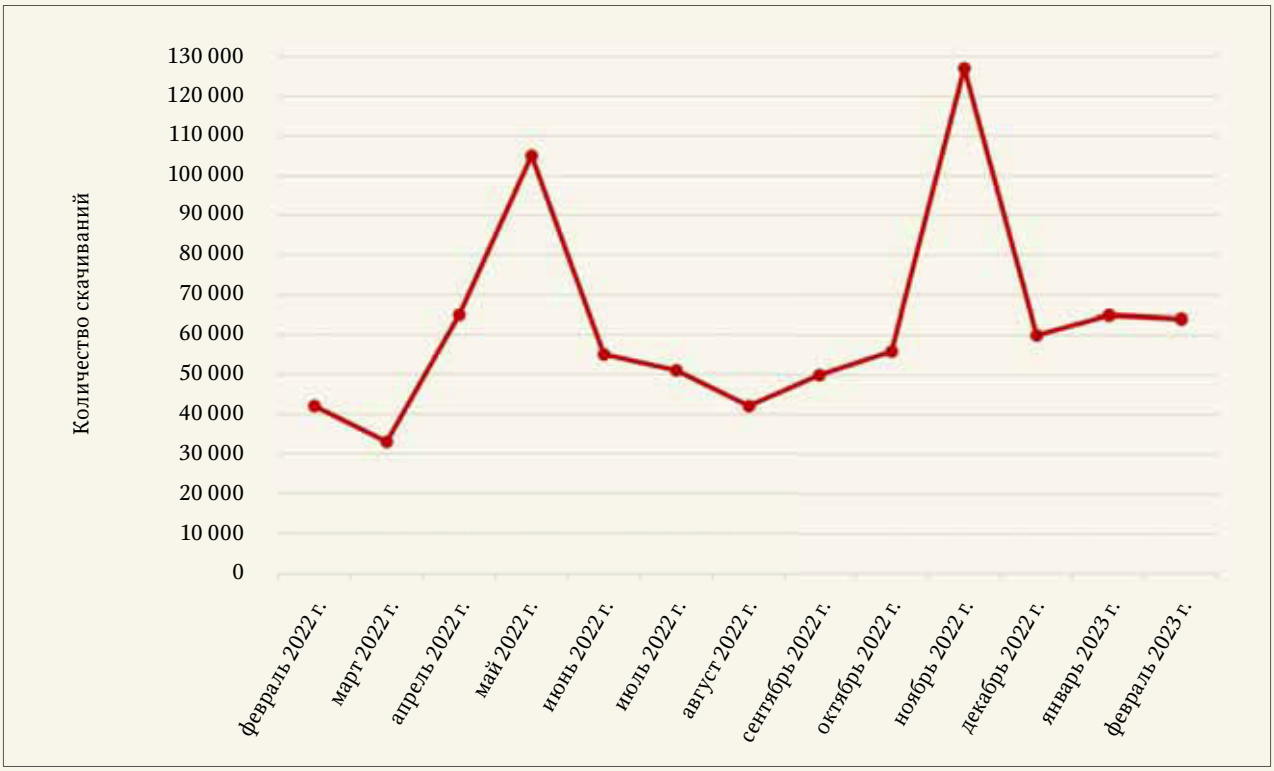

Рис. 4. Статистика скачиваний моделей ИИ, выложенных KBLab на платформе HuggingFace в открытом доступе

менным инструментам ИИ, повышающим производительность работы [20, p. 14; 40, p. 36—41].

Наиболее яркое свидетельство масштаба влияния моделей, созданных на основе библиотечных коллекций, имеет количественный характер. Согласно статистическим данным, с тех пор как они были выложены на платформе HuggingFace, было сделано более миллиона загрузок[11] (рис. 4). Эти данные, однако, требуют пояснения: речь идет не о количестве отдельных пользователей, а об общем числе раз, когда модели использовались (например, в рамках той или иной цели). Тем не менее, хотя мы не располагаем возможностью узнать дополнительные подробности об их использовании (помимо того, что можно проследить по цитированию статей сотрудников KBLab и случаев, когда они вступали в контакт с конкретными разработчиками [40, p. 20—22]), эти статистические данные все же следует рассматривать как убедительную демонстрацию масштабов проделанной работы по созданию новых инструментов ИИ. Как показали два независимых оценочных исследования первых двух лет работы KBLab, широкое распространение моделей указывает на то, что опытно-конструкторская деятельность лаборатории способствует укреплению и продвижению демократических обязательств KB как государственного учреждения [20, p. 14; 40]. Таким образом, создавая и выпуская модели ИИ с использованием данных библиотеки, KBLab предлагает новаторский и эффективный способ раскрыть ценность коллекций для широкой общественности.

Разработка ИИ на базе библиотеки как общественное благо

Основное достоинство данного подхода к разработке ИИ заключается в том, что он может одновременно обеспечить реализацию функций библиотеки и как национальной исследовательской инфраструктуры, и как гаранта демократических ценностей [9, p. 1421]. В качестве формы цифровой инфраструктуры модели с открытым доступом, созданные на основе коллекций, позволяют использовать широкий спектр приложений ИИ для шведского языка, что стало бы трудным или даже и вовсе невозможным без подобных инфраструктурных инструментов [41]. В качестве средства поощрения и обеспечения демократического развития общества существуют различные, но взаимодополняющие аспекты этих моделей, на которых стоит далее акцентировать внимание.

KB предлагает бесплатные модели, поощряя расширение реализации ИИ для шведского языка, чему способствует также особая логика архитектуры моделей. В частности, относительное распределение ресурсов между фазами предварительного обучения и тонкой настройки повышает эффективность распространения технологий: в то время как

предварительное обучение модели общего назначения, такой как KB-BERT, требует значительных вычислительных ресурсов и больших объемов данных, последующая тонкая настройка может быть выполнена с использованием лишь части данных и вычислительных ресурсов [29, p. 33]. Благодаря этому гораздо более широкий круг социальных субъектов за пределами ресурсоемких институтов, таких как университеты, получает возможность загружать модели, экспериментировать с ними и применять их в своих конкретных целях. Таким образом, работа лаборатории способствует демократизации как технологии, так и библиотечных данных [20, p. 13]. В этом смысле KBLab помогает поделиться коллекциями и ценностью, которые они теперь приобрели как данные, с новыми группами пользователей, помимо тех, кто традиционно охвачен библиотечными услугами. Иными словами, речь идет о форме финансируемого и поддерживаемого государством общего достояния [42].

Данный подход к опытно-конструкторским разработкам на базе библиотек может даже противостоять некоторым наиболее проблематичным аспектам будущего ИИ, движимым исключительно субъектами частного сектора, в частности растущему дефициту прозрачности данных. По мере того как новые технологии ИИ становятся все более совершенными и все сильнее зависят от коммерческих интересов, одновременно наблюдается тенденция к рассмотрению данных и методов обучения как коммерческой тайны, которую необходимо защищать от конкурентов. Это было продемонстрировано в недавних дискуссиях об отсутствии прозрачности в связи с выходом GPT-4 компании OpenAI. Так, исследователь по проблеме этики применения ИИ Саша Лучони (Sasha Luccioni) предположила, что с такой моделью просто невозможно заниматься наукой, учитывая отсутствие доступа к касающимся данных деталям, использованным для ее создания [43]. Такие непрозрачные методы также могут быть связаны с более широкой культурой молчания, характерной для технологической индустрии и препятствующей критическим высказываниям. Речь идет, в частности, о чрезмерной зависимости от использования огромных объемов веб-материалов неясного происхождения при обучении новых моделей, что наглядно иллюстрирует случай Тимнит Гебру (Timnit Gebru) [44; 45]. В противоположность этому, придерживаясь практики скрупулезного документирования, тщательного изучения работы на предмет репрезентативности данных и поиска более репрезентативных моделей, основанных на широте коллекций библиотеки, сотрудники лаборатории следуют прозрачному подходу к применению ИИ, что может как дополнять деятельность частных технологических компаний, так и создавать им конкуренцию [20, p. 14]. Поскольку разработка ИИ на базе библиотек в большей степени соответствует принципам открытости, а не коммерческим интересам, она может представлять собой одну из столь необходимых альтернатив чрезвычайной концентрации власти в руках нескольких крупных технологических компаний и тесно связанных с ними элитных университетов [45].

Следует отметить, что использование предметов культурного наследия и соответствующих данных при финансовой поддержке государства в качестве основы для более этичных разработок в области ИИ зависит от поиска новых источников. Несмотря на то что в лаборатории можно было продуцировать новаторские инструменты, когда они составляли часть модели BERT, темпы последних инноваций в области ИИ привели к созданию новых моделей в масштабе, который значительно усложняет эту задачу. Чтобы дать представление о произошедшем скачке, приведем следующие цифры: сеть BERT включала сотни миллионов параметров, GPT-3 насчитывает более 175 миллиардов, а GPT-4, по оценкам, — гораздо больше, чем предыдущее поколение, при этом, конечно, эта информация остается окутанной тайной. Хотя KB располагает необходимыми данными для обучения и опытом в области наук о данных для создания более крупных моделей, серьезной трудностью является поиск необходимых вычислительных ресурсов. Чтобы решить эту проблему и продолжить работу по созданию современных моделей для шведского языка, мы обратились за помощью к шведскому отделению Национального центра компетенций программы EuroCC и подали заявку на использование инфраструктуры ЕС для суперкомпьютеров в рамках инициативы EuroHPC[12]. Доступ к первому высокопроизводительному компьютеру HPC Vega (расположен на территории Словении, 240 графических процессоров), а теперь и к HPC Meluxina (расположен в Люксембурге, 800 графических процессоров) позволил лаборатории вывести свои разработки на совсем иной уровень [46]. Став первым государственным учреждением, использующим финансируемые ЕС ресурсы развития, KBLab работает над тем, чтобы сделать шведскую версию ИИ максимально открытой, прозрачной и демократичной.

Наконец, создание национальной инфраструктуры ИИ также требует поиска новых форм сотрудничества. С выпуском KB-BERT лаборатория стала ключевым игроком в развитии технологий на базе шведского языка и участником национальных и международных сетей, члены которых занимаются вопросами ИИ — от исследователей и вузовских факультетов до организаций-координаторов, государственных органов и частных компаний. Формирование новых связей и партнерских отношений с разнообразными группами помимо тех, с которыми библиотека традиционно

сотрудничала, является важным шагом в попытке добиться эффективного взаимодействия в быстро развивающемся пространстве разработок в области ИИ. Примером служит участие КВ в проекте по созданию набора контрольных показателей для оценки моделей шведского языка[13] [47] совместно с банком национальных языков Швеции в Гётеборгском университете, Шведским исследовательским институтом и Центром по развитию технологий ИИ. Работая сообща над тем, чтобы пользователям шведского ИИ было проще определить, какие модели лучше всего подходят для их целей, мы помогаем сделать последние инновации более доступными. При этом сотрудничество лаборатории с внешними субъектами также приводит к улучшению научной инфраструктуры.

Заключительные замечания

В статье обсуждалось, как создание лаборатории данных в Национальной библиотеке Швеции расширило возможности библиотеки как части цифровой исследовательской инфраструктуры. Подробно изложены практические и технические соображения, которые повлияли на создание KBLab как физического пространства, где исследователи могут получить доступ к коллекциям в невиданных ранее масштабах. Разъясняется, как цифровые коллекции библиотеки позволили лаборатории сыграть важную роль в развитии национальной инфраструктуры искусственного интеллекта для шведского языка. В заключение мы предлагаем некоторые размышления о возможностях и проблемах, с которыми сталкивается лаборатория как междисциплинарный центр разработки ИИ на базе библиотеки.

Один из ключевых аргументов в пользу создания лабораторий GLAM, в частности библиотечных лабораторий, заключается в том, что они предоставляют новые способы делиться той ценностью, которую представляют предметы культурного наследия. Как указано выше, создание подобной лаборатории может привести к эффекту снежного кома с различными положительными, хотя зачастую и непредвиденными последствиями. Так, консолидация внутреннего опыта в области изучения данных и машинного обучения открывает значительные возможности для хранящих предметы культурного наследия учреждений, которые все чаще выступают в качестве хранителей больших объемов цифровых материалов. Работая совместно со специалистами в этой предметной области (библиотекарями, сотрудниками архивов, кураторами музеев и т. д.), такие лаборатории могут сделать коллекции доступными для ученых и других пользователей, позволив им заниматься новыми направлениями научных изысканий. Применение подхода «коллекции как данные» также дает возможность внести значительный вклад в развитие ИИ, особенно для малых языков, которые не были приоритетными для крупных коммерческих игроков. Лаборатории используют высококачественные данные, основанные на собранных материалах культурного наследия, опубликованных на конкретном языке, способствуют развитию национальной инфраструктуры и налаживают партнерские связи с внешними субъектами. Все это позволяет им не только помочь сделать общественными эти данные, но одновременно заявить свое право на реализацию новой формы социально значимой деятельности. Иными словами, благодаря лабораториям GLAM создаются новаторские и неожиданные способы использования коллекций, выходящие далеко за пределы самого сектора культурного наследия.

Тем не менее существуют различные проблемы, связанные с обслуживанием подобной лаборатории и обеспечением ее будущего. Возможно, это прозвучит банально, но гораздо проще открыть лабораторию, чем закрепить ее место в качестве неотъемлемой части организации, работающей с предметами культурного наследия. Отчасти это объясняется сложностями с получением финансирования и часто наблюдающейся тенденцией к недофинансированию инфраструктуры цифровых исследований [48]. В то же время это непосредственно связано с трудностями привлечения и удержания высококвалифицированного персонала в финансируемых государством проектах в области ИИ, когда спрос на специалистов с опытом обработки данных в частном секторе постоянно растет. Более того, здесь существует также комплексная задача по интеграции этого опыта в более широкую деятельность организации: должны ли специалисты по обработке и анализу данных трудиться исключительно в рамках лаборатории, как в случае с KBLab, или же стоит распределить их компетенции по организации в целом? Как лучше всего поощрять плодотворное взаимодействие между учеными в области изучения данных и экспертами по предметной области [29, p. 45]? Требуется не только ответить на эти вопросы, но и проложить четкий маршрут в условиях стремительно меняющегося цифрового ландшафта и инноваций в области ИИ, что ведет к острой необходимости в стратегическом лидерстве и руководстве.

Создание KBLab в Национальной библиотеке Швеции было сложным, но творческим процессом, который произвел различные синергетические эффекты. Предоставление коллекций библиотеки для крупномасштабных цифровых исследований позволило создать новые инструменты ИИ, которые можно использовать как в самой библиотеке с целью повышения доступности коллекций, так и за ее пределами — для анализа данных в широком спектре контекстов. Таким образом, работа

по созданию инфраструктуры цифровых исследований и вклад в развитие национальной инфраструктуры ИИ оказались в значительной степени симбиотическими явлениями. Делясь своим отчетом об этом процессе, авторы надеются, что смогут побудить коллег на собственном опыте протестировать высказанное здесь положение о том, что будущее взаимодействие ИИ и библиотеки может быть взаимовыгодным.

Приложение 1

Правила поведения в KBLab

KBLab — это открытая дружелюбная рабочая среда, где поощряется сотрудничество, приветствуются вопросы, а стремление к критическим, открытым и независимым исследованиям является основополагающим. Эта открытость необходима для процветания новых исследовательских идей и проектов. Обязательным условием такой среды является доброжелательность друг к другу. Руководствуясь концепцией открытости, все люди, связанные с KBLab, должны вести себя в соответствии с принципами взаимного уважения и порядочности. Несоблюдение этих принципов может привести к отказу в дальнейшем доступе к KBLab. Если у вас есть какие-либо вопросы об этих нормах, просьба связаться с нами (kblabb@kb.se).

Примечания

1 https://www.westac.se/en/ (дата обращения: 15.06.2023).
2 https://liu.se/en/research/computational-text-analysis (дата обращения: 15.06.2023).
3 https://kb.se/in-english/research-collaboration/criteria-forcollaboration.html (дата обращения: 15.06.2023).
4 https://huggingface.co/KBLab/bert-base-swedish-cased (дата обращения: 15.06.2023).
5 https://huggingface.co/KBLab/wav2vec2-large-voxrex-swedish (дата обращения: 15.06.2023).
6 Данная политика распространяется на все выпущенные до сих пор модели, она сохранится и для прогностических моделей. Ситуация с генеративными моделями более сложная, поскольку существует риск использования подобных инструментов способами, противоречащими демократическим аспектам миссии библиотеки, т. е. для создания дезинформации и фейковых новостей. Исходя из этого, если лаборатория в будущем займется разработкой генеративных моделей, то в их отношении будет проводиться более взвешенная и ограничительная политика дистрибуции.
7 https://huggingface.co/KBLab (дата обращения: 15.06.2023).
8 https://kb-labb.github.io/posts/2023-01-16-sentence-transformer-20/ (дата обращения: 15.06.2023).
9 https://github.com/kb-labb (дата обращения: 15.06.2023).
10 https://www.statenssc.se/nyheter/nyhetsarkiv/2023-03-14-ai-baserad-soktjanst-ska-underlatta-remisshanteringen-i-staten (дата обращения: 15.06.2023).
11 https://github.com/kb-labb/huggingface_stats (дата обращения: 15.06.2023).
12 https://www.eurocc-access.eu/successstories/success-story-national-library-of-sweden-has-now-access-to-vega/; https://enccs.se/news/2022/10/national-library-of-sweden-accesses-meluxina/ (дата обращения: 15.06.2023).
13 https://www.ai.se/en/node/81535/superlim (дата обращения: 15.06.2023).

Список литературы

1. *Hoy M.B.* Big Data: An Introduction for Librarians // Medical Reference Services Quarterly. 2014. Vol. 33, № 3. P. 320—326. DOI: 10.1080/02763869.2014.925709.
2. *Cukier K., Mayer-Schoenberger V.* The Rise of Big Data: How It's Changing the Way We Think About the World // Foreign Affairs. 2013. Vol. 92, № 3. P. 28—40.
3. *Bingham N.J., Byrne H.* Archival Strategies for Contemporary Collecting in a World of Big Data: Challenges and Opportunities with Curating the UK Web Archive // Big Data & Society. 2021. Vol. 8, № 1. DOI: 10.1177/2053951721990409.
4. *Ames S., Lewis S.* Disrupting the Library: Digital Scholarship and Big Data at the National Library of Scotland // Big Data & Society. 2020. Vol. 7, № 2. DOI: 10.1177/2053951720970576.
5. *Underwood T.* Distant Horizons: Digital Evidence and Literary Change. Chicago, Illinois : The University of Chicago Press, 2019. 200 p.
6. *Jasanoff S., Kim S.* Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea // Minerva. 2009. Vol. 47, № 2. P. 119—146.
7. *Padilla T.* Humanities Data in the Library: Integrity, Form, Access // D-Lib Magazine. 2016. Vol. 22, № 3/4. DOI: 10.1045/march2016-padilla.
8. *Mahey M., Al-Abdulla A., Ames S., Bray P., Candela G., Chambers S., Derven C., Dobreva-McPherson M., Gasser K., Karner S., Kokegei K., Laursen D., Potter A., Straube A., Wagner S.-C., Wilms L.* Open a GLAM Lab. Digital Cultural Heritage Innovation Labs. Doha, Qatar : Book Sprint, 2019.
9. SFS 2008:1421 Förordning med instruktion för Kungl. biblioteket. URL: https://www.lagboken.se/Lagboken/start/sfs/sfs/2008/1400-1499/d_207271-sfs-2008_1421-forordning-med-instruktion-for-kungl.-biblioteket (дата обращения: 22.09.2023).
10. *Konstenius G.* Plikten under lupp! En studie av pliktlagstiftningens roll, utformning och relevans i förhållande till medielandskapets utveckling [Eng. Legal Deposit in Focus! A Study of the Role, Design and Relevance of Legal Deposit Legislation in Relation to the Development of the Media Landscape]. Stockholm : Kungl. Biblioteket, 2017. URL: https://libris.kb.se/bib/21946684 (дата обращения: 22.09.2023).
11. *Padilla T.* Responsible Operations: Data Science, Machine Learning, and AI in Libraries. Dublin, Ohio, USA : Online Computer Library Center, Inc., 2019. 38 p.
12. *Padilla T., Allen L., Frost H., Potvin S., Russey Roke E., Varner S.* Final Report — Always Already Computational:

Collections as Data // Zenodo, 2019. 180 p. URL: https://zenodo.org/record/3152935 (дата обращения: 22.09.2023).
13. *Traub M.C., van Ossenbruggen J., Hardman L.* Impact Analysis of OCR Quality on Research Tasks in Digital Archives // Research and Advanced Technology for Digital Libraries / ed. by S. Kapidakis, C. Mazurek, C. and M. Werla. Cham : Springer, 2015. P. 252—263.
14. *Rekathati F.* A Multimodal Approach to Advertisement Classification in Digitized Newspapers // The KBLab Blog. March 28, 2021. URL: https://kb-labb.github.io/posts/2021-03-28-ad-classification/ (дата обращения: 22.09.2023).
15. "Raw Data" is an Oxymoron / ed. by L. Gitelman. Cambridge, Mass. : The MIT Press, 2013. 192 p.
16. *Snickars P.* Datalabb på KB: En förstudie. [Eng. Data Lab at KB: A Pre-study]. Stockholm : Kungl. Biblioteket, 2018.
17. *Law J.* After Method: Mess in Social Science Research. London : Routledge, 2004. 188 p.
18. *Malmsten M.* KB Data Lab // Github. URL: https://github.com/Kungbib/kblab (дата обращения: 22.09.2023).
19. *Malmsten M.* Exposing Library Data as Linked Data // IFLA Satellite Preconference Sponsored by the Information Technology Section, 2009.
20. *Fridlund M.* Utvärdering av KB-labb. [Eng. Evaluation of KBLab]. Göteborgs Universitet, Centrum för digital humaniora, 2021. 38 p. URL: https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-97 (дата обращения: 22.09.2023).
21. *Rekathati F.* Curating News Sections in a Historical Swedish News Corpus : Independent Master's Thesis. Linköping University, Department of Computer and Information Science, 2020. URL: https://liu.diva-portal.org/smash/record.jsf?pid=diva2%3A1438672&dswid=1966 (дата обращения: 22.09.2023).
22. *Henning G.* News Article Segmentation Using Multimodal Input: Using Mask R-Cnn and Sentence Transformers : Independent Master's Thesis. Stockholm : Kungliga Tekniska högskolan, School of Electrical Engineering and Computer Science, 2022. 78 p.
23. *Kemman M.* Trading Zones of Digital History. De Gruyter Oldenbourg, 2021. 182 p.
24. *Fano E., Haffenden C.* Digital humaniora eller humanistisk datavetenskap? [Eng. Digital Humanities or Humanistic Computer Science?] // Kungl. Biblioteket. Samlingsbloggen. 14 april 2022. URL: https://www.kb.se/hitta-och-bestall/samlingsbloggen/blogginlagg/2022-04-14-digital-humaniora-eller-humanistisk-datavetenskap.html (дата обращения: 22.09.2023).
25. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 : Long and Short Papers. Minneapolis, Minnesota, 2019. P. 4171—4186. URL: https://aclanthology.org/N19-1423/ (дата обращения: 22.09.2023).
26. *Virtanen A., Kanerva J., Ilo R., Luoma J., Luotolahti J., Salakoski T., Ginter F., Pyysalo S.* Multilingual Is Not Enough: BERT for Finnish // arXiv:1912.07076. DOI: 10.48550/arXiv.1912.07076.
27. *Martin L., Muller B., Súarez P.J.O., Dupont Y., Romary L., de la Clergerie E.V., Seddah D., Sagot B.* CamemBERT: a Tasty French Language Model // arXiv:1911.03894. DOI: 10.48550/arXiv.1911.03894.
28. *Kummervold P.E., de la Rosa J., Wetjen F., Brygfjeld S.A.* Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model // Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa). Reykjavik, Iceland (Online) : Linköping University Electronic Press, 2021. P. 20—29. URL: https://aclanthology.org/2021.nodalida-main.3/ (дата обращения: 22.09.2023).
29. *Haffenden C., Fano E., Malmsten M., Börjeson L.* Making and Using AI in the Library: Creating a BERT Model at the National Library of Sweden // College & Research Libraries. 2023. Vol. 84, № 1. P. 30—48. URL: https://crl.acrl.org/index.php/crl/article/view/25748/33669 (дата обращения: 22.09.2023).
30. *Radford A., Narasimhan K., Salimans T., Sutskever I.* Improving Language Understanding by Generative Pre-Training. [pre-print], 2018.
31. *Malmsten M., Börjeson L., Haffenden C.* Playing with Words at the National Library of Sweden: Making a Swedish BERT // arXiv:2007.01658. DOI: 10.48550/arXiv.2007.01658.
32. *Malmsten M., Haffenden C., Börjeson L.* Hearing Voices at the National Library: A Speech Corpus and Acoustic Model for the Swedish Language // arXiv:2205.03026. DOI: 10.48550/arXiv.2205.03026.
33. *Baevski A., Zhou H., Mohamed A., Auli M.* wav2vec 2.0: A Framework for SelfSupervised Learning of Speech Representations // arXiv:2006.11477. DOI: 10.48550/arXiv.2006.11477.
34. *Carlsson F., Eisen P., Rekathati F., Sahlgren M.* Cross-lingual and Multilingual CLIP // Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022. URL: https://aclanthology.org/2022.lrec-1.739 (дата обращения: 22.09.2023).
35. *Estmark A.* Text Block Prediction and Article Reconstruction Using BERT : Independent Master's Thesis. Uppsala University, Department of Statistics, 2021. 41 p.
36. *Fano E., Haffenden C.* BERTopic for Swedish: Topic Modeling Made Easier via KBBERT // The KBLab Blog. June 14, 2022. URL: https://kb-labb.github.io/posts/2022-06-14-bertopic/ (дата обращения: 22.09.2023).
37. *Dwibedi C., Mellergård E., Gyllensten A.C., Nilsson K., Axelsson A.S., Bäckman M., Sahlgren M., Friend S.H., Persson S., Franzén S., Abrahamsson B., Steen Carlsson K., Rosengren A.H.* Effect of Self-Managed Lifestyle Treatment on Glycemic Control in Patients with Type 2 Diabetes // NPJ Digital Medicine. 2022. Vol. 5, № 1. Art. 60. URL: https://portal.research.lu.se/en/publications/effect-of-self-managed-lifestyle-treatment-on-glycemic-control-in (дата обращения: 22.09.2023).
38. *Jerdhaf O., Santini M., Lundberg M., Karlsson A.* Implant Terms: Focused Terminology Extraction with Swedish BERT — Preliminary Results // Eighth Swedish Language Technology Conference. 2020. URL: http://

www.diva-portal.org/smash/record.jsf?pid=diva2%3A1526369&dswid=5926 (дата обращения: 22.09.2023).
39. *Avram A., Pais V., Tufis D.* PyEuroVoc: A Tool for Multilingual Legal Document Classification with EuroVoc Descriptors // Proceedings of the International Conference on Recent Advances in Natural Language Processing. 2021. URL: https://aclanthology.org/2021.ranlp-1.12/ (дата обращения: 22.09.2023).
40. *Juhlin M.* De samhällsekonomiska effekterna kopplat till Kungliga bibliotekets AIbaserade språkmodeller. [Eng. The Socioeconomic Effects of the National Library of Sweden's AI Language Models]. Policy Impact report. 2022. URL: https://www.kb.se/samverkan-och-utveckling/kb-labb.html (дата обращения: 22.09.2023).
41. *Edwards P.N., Bowker G.C., Jackson S.J., Williams R.* Introduction: An Agenda for Infrastructure Studies // Journal of the Association for Information Systems. 2009. Vol. 10, № 5. P. 364–374.
42. *Harvey D.* The Future of the Commons // Radical History Review. 2009. № 109. P. 101–107. DOI: 10.1215/01636545-2010-017.
43. *Sanderson K.* GPT-4 Is Here: What Scientists Think // Nature. 2023. Vol. 615, № 7954. Art. 773. DOI: 10.1038/d41586-023-00816-5.
44. *Bender E.M., Gebru T., McMillan-Major A., Shmitchell S.* On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // FAccT '21: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. 2021. P. 610–623. DOI: 10.1145/3442188.3445922.
45. *Gebru T.* For Truly Ethical AI, Its Research Must Be Independent from Big Tech // The Guardian. Dec. 6, 2021.
46. *Kurtz R., Öhman J.* SUCX 3.0 // The KBLab Blog. Feb. 7, 2022. URL: https://kb-labb.github.io/posts/2022-02-07-sucx3_ner/ (дата обращения: 22.09.2023).
47. *Kurtz R.* Evaluating Swedish Language Models. The KBLab Blog. March 16, 2022. URL: https://kb-labb.github.io/posts/2022-03-16-evaluating-swedish-language-models/ (дата обращения: 22.09.2023).
48. *Knowles R., Mateen B.A., Yehudi Y.* We Need to Talk About the Lack of Investment in Digital Research Infrastructure // Nature Computational Science. 2021. Vol. 1, № 3. P. 169–171. DOI: 10.1038/s43588-021-00048-5.

Перевод **Марии Федотовой**,
Российская государственная библиотека

Анонс

Вебинар «Концепция MACK: Развитие местной библиотечной практики для эффективной деятельности по борьбе с изменением климата»

Дата:	19 января 2024 г.
Место проведения:	онлайн
Время начала мероприятия:	9:00 по тихоокеанскому времени (17:00 по Гринвичу)
Регистрация:	требуется (бесплатно)
Мероприятие будет записано:	да
Докладчик:	Дэн Хэкборн (Dan Hackborn)

В рамках цикла вебинаров Секции ИФЛА по вопросам окружающей среды, устойчивого развития и библиотек (ENSULIB) Дэн Хэкборн представит доклад на тему «Концепция MACK: Развитие местной библиотечной практики для эффективной деятельности по борьбе с изменением климата». Концептуализация эффективных действий по борьбе с изменением климата в рамках отдельного учреждения может оказаться сложной задачей в связи с глобальным характером антропогенного изменения климата. На семинаре участникам расскажут о полезном инструменте — концептуальной модели «Митигация — Адаптация — Сообщество — Знания» (MACK), который позволяет поместить действия по борьбе с изменением климата в контекст местной библиотеки. На конкретном примере будет показано, как эта модель может быть использована руководителями библиотек для уточнения и анализа уникальных возможностей своего учреждения в области борьбы с изменением климата с учетом потребностей сообщества.

Организаторы цикла:
- Бет Филар Уильямс (Beth Filar Williams), beth.filar-williams@oregonstate.edu
- Вивьен Бёрд (Vivienne Byrd), vbyrd@lapl.org
- Антония Мокатта (Antonia Mocatta), antonia.mocatta@sydney.edu.au
- Присцилла Пун (Priscilla Pun), nipun@um.edu.mo
- Харри Сахавирта (Harri Sahavirta), harri.sahavirta@hel.fi
- Петра Хауке (Petra Hauke), petra.hauke@hu-berlin.de