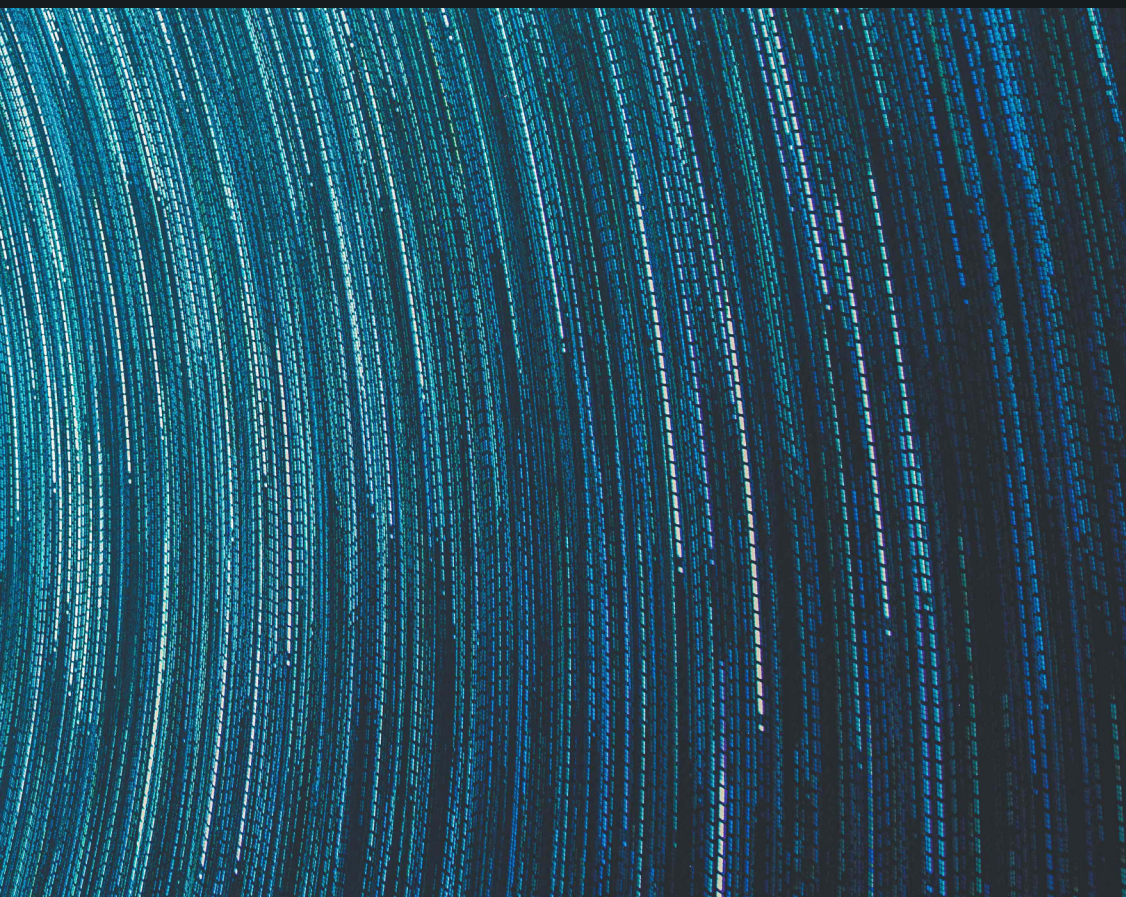
СИБИРСКИЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

ЦИФРОВЫЕ ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ



Министерство науки и высшего образования Российской Федерации
Сибирский федеральный университет

ЦИФРОВЫЕ ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ

Монография

Красноярск
СФУ
2023

УДК 009:004.0
ББК 71.034+32.97
Ц752

Авторы:
А.Б. Антопольский, А.А. Бонч-Осмоловская, Л.И. Бородкин, А.Ю. Володин, Д.А. Гагарина, Е.С. Гришин, И.А. Кижнер, Б.В. Орехов, М.В. Румянцев, А.В. Сметанин

Рецензенты:
С. А. Баканов, доктор исторических наук, заведующий кафедрой истории России и зарубежных стран историко-филологического факультета Челябинского государственного университета;
В. Н. Владимиров, доктор исторических наук, профессор кафедры отечественной истории Института истории и международных отношений Алтайского государственного университета, вице-президент Ассоциации «История и компьютер»;
Е. А. Пастернак, кандидат филологических наук, научный сотрудник Института мировой культуры Московского государственного университета им. М. В. Ломоносова

Ц752 **Цифровые гуманитарные исследования** : монография / А.Б. Антопольский, А.А. Бонч-Осмоловская, Л.И. Бородкин [и др.]. — Красноярск : Сиб. федер. ун-т, 2023. — 272 с.
ISBN 978-5-7638-4876-2

Впервые на русском языке комплексно рассмотрено актуальное междисциплинарное направление — цифровые гуманитарные исследования, или digital humanities. Приведены примеры (само)определения направления, дан их обзор. «Цифровой поворот» в гуманитарных исследованиях и масштабные проекты оцифровки историко-культурного наследия описаны в контексте датафикации и вызовов больших данных и машинного обучения. Особое внимание уделено современным подходам к компьютерному анализу текстов и культуромике, направлению исследований культуры и языка с помощью больших текстовых данных. Представлена широкая палитра цифровых подходов, призванных находить решения насущных гуманитарных исследовательских задач: от базы данных к сетевому анализу, от геоинформационных систем к виртуальным реконструкциям и дополненной реальности. Происходящие процессы рассмотрены в связи со становлением сложной и противоречивой информационной инфраструктуры цифровых гуманитарных исследований.

Будет интересна широкому кругу гуманитариев — историкам, филологам, философам, культурологам — и всем сочувствующим и сопереживающим цифровой трансформации современной культуры.

Электронный вариант издания см.:
http://catalog.sfu-kras.ru

УДК 009:004.0
ББК 71.034+32.97

ISBN 978-5-7638-4876-2

Оглавление

Предисловие 4

Глава 1. Digital humanities: (само)определение,
обзор направлений 5

Глава 2. Данные в цифровых гуманитарных исследованиях 21

Глава 3. Культурное наследие и цифровые коллекции данных 39

Глава 4. Культуромика: исследование культуры и языка
с помощью больших текстовых данных 57

Глава 5. Базы данных: модели, структуры, связанные данные 100

Глава 6. Компьютерный анализ текста 120

Глава 7. Геоинформационные системы: подходы,
методики, данные 158

Глава 8. 3D-моделирование, виртуальные реконструкции
и VR/AR/MR-технологии в задачах сохранения
культурного наследия 186

Глава 9. Сетевой анализ данных (social network analysis, SNA):
подходы и технологии 221

Глава 10. Информационная инфраструктура цифровых
гуманитарных исследований 244

Послесловие 264

Информация об авторах 267

Предисловие

Идея коллективного труда, который вы держите в руках, появилась во время подготовки онлайн-курса «Введение в цифровые гуманитарные исследования» (https://openedu.ru/course/sfu/IDH/) на базе производственно-продюсерского центра Сибирского федерального университета.

Цифровые гуманитарные науки (или digital humanities) стали важным направлением развития современных методических подходов к решению исторических, лингвистических, культурологических, философских проблем.

Однако все еще ощущается недостаток в литературе на русском языке, которая могла бы поспособствовать вхождению желающим в такую сложную, но захватывающую проблематику.

В издательстве Сибирского федерального университета увидела свет хрестоматия «Цифровые гуманитарные науки», переводное издание, отразившее сложные перипетии становления и самоопределения направления. Хрестоматия вошла в программы большинства смежных дисциплин и получила признание у преподавателей. Но исследования не стоят на месте, в России активно развиваются разные методологические направления цифровых гуманитарных исследований, настал удачный момент соединить исследовательские наработки с опытом преподавания дисциплин цифрового цикла.

Под одной обложкой вы найдете главы о данных и базах данных, о культуромике и анализе текстов, о географических информационных системах и сетевом анализе, о трехмерном моделировании и об инфраструктурах современной цифровой гуманитарной науки.

Не все аспекты богатого исследовательского поля цифровой гуманитаристики удалось охватить в этом первом коллективном опыте, но есть надежда, что коллективная монография станет живой — обновляющейся и дополняющейся от издания к изданию. И наши труды будут способствовать развитию и процветанию цифровой гуманитарной науки в России.

Максим Румянцев,
ректор СФУ

Глава 1

Digital humanities: (само)определение, обзор направлений

(*А. Ю. Володин, Б. В. Орехов*)

Digital humanities — направление исследований, которое за два десятилетия завоевало свое место в научной повестке гуманитарных междисциплинарных компьютеризированных изысканий. Компьютеризация в гуманитарных науках, в принципе, имеет давние традиции: к помощи компьютерной техники ученые-гуманитарии начали обращаться еще в эпоху больших вычислительных машин. Но настоящий «цифровой поворот» в гуманитарных науках начался после микрокомпьютерной революции, с развитием вычислительных мощностей и персонализацией компьютерных систем, позволяющих производить сложные подсчеты в домашних условиях, не только создавать сложные виртуальные реконструкции, но и представлять их в электронной среде с помощью многообразных средств Всемирной паутины. За последние десятилетия существенно снизился порог входа в тематику цифровых методов, все больше ученых вовлекается в орбиту количественных исследований, что приводит к значительным институциональным и концептуальным изменениям в смежных гуманитарных дисциплинах.

Термину digital humanities не повезло с переводом на русский язык. Проблема кроется в отсутствии в русскоязычной традиции удобного эквивалента для слова humanities, которое трудно перевести кратко. Наиболее адекватным является длинный перевод из двух слов «гуманитарные науки», в реальном употреблении он как будто требует сокращения, и это сокращение находится в неудачном слове «гуманитаристика».

Неудачность этого слова в том, что стилистически оно не нейтрально, то есть создает ненужный сниженный эффект, давно

замеченный лингвистами для слов с формантом -истика (шагистика, ерундистика)[1]. В более выгодном положении оказываются давно заимствованные слова «журналистика», «логистика», «лингвистика», но, чтобы оказаться в этом ряду, все еще имеющей ореол новизны «гуманитаристике» придется проделать длинный и сложный путь адаптации в русском языке. Отдельно от определения «цифровой» это слово в русском научном языке почти не употреблялось, поэтому и не воспринимается как привычный нейтральный термин.

Новизна здесь не случайна: digital humanities появились в научном поле как самостоятельная сфера всего несколько десятилетий назад, поэтому их границы, объект, методы и ценности не успели приобрести отчетливых очертаний. Процесс оформления в дисциплину продолжается на наших глазах, поэтому любые обобщения, сделанные сейчас, будут носить предварительный характер.

Если максимально упростить схему, то digital humanities являются продолжением соответствующих областей гуманитарного знания. Цифровые гуманитарные науки распадаются на цифровое литературоведение, электронные публикации, цифровое искусствоведение (включая музееведение), цифровое киноведение. Но важным направлением внутри digital humanities является и цифровая история, несмотря на то, что статус истории как гуманитарной, а не социальной науки остается дискуссионным[2]. В этом смысле компонент «гуманитарный» в составе термина «цифровые гуманитарные науки» оказывается и обманчивым, и точным.

Точен он потому, что в отечественной традиции история обычно локализуется внутри гуманитарного поля. Если смотреть на это через призму институционализации, то исторические дисциплины изучаются внутри гуманитарного блока учебных предметов, кафедры истории размещены на гуманитарных факультетах, а исторические факультеты являются частью гуманитарных вузов (например, Российский государственный гуманитарный университет возник на базе Историко-архивного института). Еще в середине XX века привычным для советского вуза было наличие общего историко-филологического факультета.

---

[1] Пацюкова О. А. Переразложение и закономерности развития протяженных аффиксов в русском языке: дис. … д-ра филол. наук. Нижний Новгород, 2014. С. 186.

[2] Валлерстайн И. Миросистемный анализ. Введение. М.: УРСС: ЛЕНАНД, 2018. С. 60–65.

Обманчив он потому, что не дает ясного понимания специфики digital humanities. Как выглядела бы эта область без истории? Если бы digital humanities состояли только из цифрового литературоведения (computational literary studies), цифрового искусствоведения и других подобных дисциплин, то можно было бы сравнительно четко очертить круг исследовательских интересов, составляющих ядро цифровых гуманитарных наук. Этот круг включал бы прежде всего семиотические объекты второго порядка или целые вторичные семиотические системы.

Семиотика — это наука о знаковых системах. Эти системы позволяют людям обмениваться информацией. Например, знаковую систему составляют сигналы светофоров, дорожная разметка, азбука Морзе, математические формулы, музыкальная нотация. Красный свет светофора — это знак, у него есть значение: команда остановиться. Буква Σ в математической формуле — это знак, у него есть значение: сумма. Но самой сложной и развитой знаковой системой в распоряжении человека является естественный язык: русский, английский, амхарский, рутульский и т.д. Описанием того, как устроен и как функционирует язык, занимается наука лингвистика (стереотипное представление, будто бы лингвистика — это изучение иностранных языков, следует считать распространенным заблуждением; кстати, заблуждение, по всей видимости, и то, что лингвистика относится к гуманитарным наукам, она гораздо ближе естественным, вроде биологии). Лингвистика создала сложный терминологический аппарат для описания слов, их грамматических характеристик и семантики, способов сочетания этих слов между собой в предложении и тексте. Выше мы использовали как раз лингвистические термины: «формант», «стилистическая нейтральность». Пользуясь лингвистическим знанием, можно не только описать, какие слова и как расставлены в тексте, но и объяснить, почему высказывание устроено именно так.

Базовым основанием лингвистики как науки является представление о языке как о всеобщем и самостоятельном инструменте общения. Иными словами, язык знают все его носители, он не принадлежит кому-то одному (поэтому всеобщий), а еще то, что мы знаем о языке, невозможно вывести из наших знаний о человеческой биологии или психологии (поэтому самостоятельный). Это довольно важная причина считать языкознание самостоятельной наукой: ни одна другая научная дисциплина не способна описать согласовательные классы слов, актантную деривацию или временной дейксис.

Язык — это семиотическая система первого порядка. С помощью содержащихся в языке знаков мы передаем друг другу информацию. Но, используя язык, люди стали создавать такие объекты, описать и объяснить которые лингвистика уже не может и не стремится. Речь о художественной литературе. В литературных произведениях появились свои знаки, которые так или иначе имеют языковое выражение, но которые невозможно свести к языку и ограничиться лингвистическим описанием.

К таким знакам, например, можно отнести кольцевую композицию, лирические отступления в нарративном тексте, систему персонажей. Лирическое отступление в повествовательном по своей сути романе в стихах «Евгений Онегин» — это знак, он указывает читателю на игру Пушкина с традицией лиро-эпических поэм Байрона. Кольцевая композиция рассказа Набокова «Круг» — это знак, он актуализирует форму организации текста, выдвигает ее на первый план в восприятии произведения по отношению к содержанию. Появление в детективном сюжете опытного разгадывателя загадок и его наивного, но верного и преданного спутника, как в романе «Имя Розы» Умберто Эко, — это знак, он вызывает в памяти ассоциации с классическими рассказами о Шерлоке Холмсе.

Знаковая природа таких элементов текста очевидна, она использует язык как материал, но недоступна для описания с помощью терминологического аппарата лингвистики, потому что названные (а также множество не упомянутых здесь) элементы созданы в иной семиотической системе. Это уже не система языка, а надстроенная над ней система литературы, имеющая единицы, правила и закономерности.

Такими семиотическими системами второго порядка, надстроенными над первичными, но не сводимыми к последним, и занимается гуманитарная наука. Сигналы светофора — это семиотика первого порядка, с ее помощью передаются сигналы участникам дорожного движения. Светофор, изображенный на картине или на художественной фотографии, встраивается в систему второго порядка, участвует в композиции кадра, определяет цветность и мотивное наполнение изображения. К регулированию действий пешеходов и автомобилистов все это отношения уже не имеет.

Но история в этом смысле не похожа на гуманитарные науки, ее предметом не являются семиотические системы ни первого, ни второго порядка. Разумеется, историк может привлекать к своим

изысканиям текстовые источники — так же, как это делает лингвист или литературовед, но делает он это с другими целями.

Историческое знание принципиально опосредовано объекту своего исследования и основывается на реконструкции. Исследователь восстанавливает прошлое на основе исторических источников, «остатков» и «преданий» старины. Можно назвать такую реконструкцию информационным моделированием прошлого. Как верно подметил Ю. М. Лотман, «историк с самого начала попадает в странное положение: в других науках исследователь начинает с фактов, историк получает факты как итог определенного анализа, а не в качестве его исходной точки»[1]. Историки традиционно чутки к различиям в средствах передачи информации (как в знаменитой формуле М. Маклюэна — «The medium is the message» ‘то, что передает сообщение, само по себе является сообщением’). Доказательством тому служит долгая дискуссия о классификации исторических источников и разнообразие специальных исторических дисциплин, ориентированных на конкретные виды исторических источников.

То общее, что интересует и историков, и представителей типичных гуманитарных наук, лежит в особой плоскости, а именно в плоскости противопоставления номотетических и идиографических дисциплин.

Слово «номотетический» восходит к греческому *nomos* ‘закон’, этим термином объединяются науки, которые работают с повторяющимися явлениями, а их конечной целью является описание законов и закономерностей. Хрестоматийные примеры — из классической механики вроде ускорения свободного падения. Нахождение таких законов и их математическое описание является сверхзадачей естественных наук: физики, химии, физиологии и подобных.

Им противостоят идиографические науки, которые сосредоточены на уникальных явлениях, к которым относятся исторические события и события истории литературы и искусства. Приход к власти Елизаветы Петровны, роман «Обломов» или стихотворение «Письмо матери» рассматриваются как единственные в своем роде, и в оптике этих дисциплин не могут стать частным случаем проявления какого-то закона. К идиографическим относятся и история, и гуманитарные науки.

[1] Лотман Ю. М. Изъявление Господне или азартная игра? (Закономерное и случайное в историческом процессе) // Ю. М. Лотман и тартуско-московская семиотическая школа. М.: Гнозис, 1994. С. 353–363.

Именно в этом кроется нетипичность цифрового подхода к гуманитарному материалу. Количественные методы хорошо зарекомендовали себя там, где анализ повторяющихся фактов помогает открывать и описывать закономерности, подтверждаемые новыми наблюдениями. То, что является ценным для идиографических дисциплин, не повторяется и не может быть подтверждено.

Чтобы перейти от традиционных идиографических описаний к применению количественных методов, необходимо перестроить сам взгляд на материал таким образом, чтобы он позволял видеть общее в уникальном. Например, не ставя под сомнение единственность в своем роде каждого отдельного стихотворения, ученый может сосредоточиться на некотором формальном приеме, истории его применения в рамках определенной художественной традиции. В таком случае попавший в фокус внимания исследователя прием уже может иметь количественное измерение, которое может быть описано как закономерность. Так действуют применяющие квантитативные методы стиховеды, подсчитывающие частотность использования поэтических размеров: «Обследуются не типы, а массы, не единичные явления, а общие состояния. Рассматриваются многие тысячи стихов данного размера, рассматриваются в различных разрезах, и — “большие числа” торжествуют над случайностью, выстраивая закон»[1].

Такой подход, естественно, вызывает закономерный скепсис в глазах традиционных филологов и искусствоведов. Подсчет частотности вступает для них в противоречие с привычными формами работы с материалом. Поэтому неверно было бы разделять историю гуманитарных наук на доцифровой и цифровой периоды: одновременно с активным использованием компьютерных методик существуют и более привычные подходы.

Тем не менее вряд ли можно найти гуманитарную специальность, которую бы так или иначе не затронул «цифровой поворот». Любое гуманитарное исследование сегодня основано на спонтанной или систематической, выборочной или сплошной оцифровке текстов, документов, изображений или каких-то объектов историко-культурного наследия, что делает эти объекты более доступными для исследователя, то есть снижает порог входа для ученого. Если в доцифровую эру на пути у филолога, историка, искусствоведа могла

[1] Шенгели Г. А. Об исследовании узбекского стиха // Научная мысль. 1930. № 1. С. 29.

стоять недоступность текста, живописного полотна или документа, то в современной ситуации наличие цифровой копии вовлекает в исследовательский процесс все большее число специалистов. Так, академик Е. Э. Бертельс, составляя историю персидско-таджикской литературы, обозначал масштаб проблемы: «ссылаться на книгу, которая имеется только, скажем, в библиотеке Института востоковедения в Ленинграде и которую нельзя найти даже и в Москве, или на рукопись, находящуюся в частной библиотеке, было бы просто недобросовестно»[1]. Оцифровка стала одной из важных ежедневных практик ремесла гуманитария. В этой связи возникает широкий спектр вопросов, в чем преимущества и недостатки наступления цифровой эры в гуманитарных исследованиях. Такие вопросы оказываются главными в весьма обширной литературе, посвященной проблемам определения, самоопределения и развития междисциплинарного направления digital humanities.

В каком-то смысле возможности электронной публикации и сетевого доступа начинают играть роль «дополненной реальности», когда классические формы научного творчества (статьи, монографии) дополняются электронными ресурсами, содержащими цифровые приложения, часто имеющие самостоятельное научное значение. Следует отметить, что в среде специалистов наблюдается относительный консенсус о том, что цифровые гуманитарные науки предполагают не только использование компьютера как исследовательского инструмента, но и расширение цифрового историко-культурного наследия путем публикации электронных ресурсов, реконструкций и визуализаций. Без таких публикаций исследование может быть компьютеризированным, но не может относиться к направлению DH.

По той же причине для истории использование компьютерных методов оказывается особенно актуальным. Сегодня можно по-новому взглянуть на знаменитую максиму Э. Ле Руа Ладюри — «историк будущего будет программистом или его не будет вовсе» (“l’historien de demain sera programmeur ou il ne sera plus”; Le nouvel observateur, 1968). Историки, бесспорно, стали пользователями (а иногда и создателями) весьма разнообразного программного обеспечения. Начали реализовываться давние мечты об историко-ориентированном программном обеспечении. К таким разработкам

[1] Бертельс Е. Э. Избранные труды. Т. 1. История персидско-таджикской литературы. М.: Изд-во восточной литературы, 1960. С. 28.

можно отнести уже приобретшие мировую известность продукты Центра истории и новых медиа имени Р. Розенцвейга: программа Zotero позволяет сохранять и управлять найденными онлайн научными материалами; Omeka предназначена для создания специализированных электронных ресурсов — электронных коллекций и онлайн-выставок; Scripto создан для облегчения совместной работы по расшифровке и установлению текстов по электронным копиям архивных документов. Именно по этой причине рассмотрение новых возможностей работы с источниками информации определяет профессиональные аспекты исторического исследования в цифровую эпоху. Историки сосредоточились на изучении исторических источников, представлении исторических сведений в формате баз данных, оцифровке и электронной публикации свидетельств прошлого[1], а вслед за оцифровкой — на моделировании исторических процессов и объектов в самом широком смысле этого понятия: от математических моделей поведения до трехмерных моделей объектов прошлого[2].

Цифровая история (хотя такой перевод digital history и вызывает критику) сегодня часто рассматривается как область активной разработки инструментов обработки и исследования исторических источников для их адекватного представления в современных медиаформатах (в последние годы преимущественно онлайн). Термин digital history приобрел права гражданства в 1997 году, когда американские исследователи Э. Айерс и У. Томас основали Вирджинский центр цифровой истории (Virginia Center for Digital History, VCDH) при университете Вирджинии, хотя один из пионеров разработок в этой области Р. Розенцвейг еще в 1994 году открыл Центр истории и новых медиа (Center for History and New Media, CHNM) в университете Дж. Мейсона. Первые работы, посвященные осмыслению цифровой истории, были написаны на рубеже XX–XXI веков, в частности, можно отметить полемическую статью Э. Айерса «Прошлое и будущее цифровой истории» и фундаментальную монографию Д. Коэна и Р. Розенцвейга «Цифровая история: руководство по сбору, сохранению и представлению прошлого во Всемирной паутине»[3].

---

[1] Rosenzweig R., Grafton A. Clio Wired: The Future of the Past in the Digital Age. Columbia University Press, 2011.

[2] Бородкин Л. И. Моделирование исторических процессов: от реконструкции реальности к анализу альтернатив. СПб.: Алетейя, 2016.

[3] Ayers Edward. L. 1999. The Pasts and Futures of Digital History (online essay). URL: http://www.vcdh.virginia.edu/PastsFutures.html (дата обращения: 26.08.2023);

В коллективной монографии «История в цифровую эпоху» под редакцией Т. Веллер сформулирован ключевой для историков подход к цифровой эпохе — «цифра» затрагивает всех, кто профессионально изучает историю, при этом из этого не следует, что историк обязательно должен становиться компьютерным гуру или разбираться в языках программирования. Главное, чтобы информационные технологии и цифровые достижения, напрямую касающиеся профессиональных нужд историков, не оставались закрытым самодостаточным полем исследований отдельных специалистов (как частично получилось в случае с «количественной историей»), не становились маргинальными вопросами хотя бы по той причине, что по большей части цифровая эпоха касается основ методологии и методики исторического исследования[1].

Кроме того, сегодня заметен определенный историографический переход, состоявшийся с развитием средств компьютерной визуализации и сетевых технологий. Данный переход можно условно датировать серединой 2000-х годов, когда постепенно стало происходить терминологическое изменение: от исторического или гуманитарного компьютинга (humanities computing, history and computing) к цифровым гуманитарным наукам или цифровой истории (digital humanities, digital history). Перемена названия означала постепенное изменение статуса — от технической поддержки к интеллектуальному прорыву со своими профессиональными практиками, научными стандартами и теоретическими построениями. Во многом переход от «измерительных» возможностей компьютерных технологий к реконструкционным и презентационным связан с освоением интернет-технологий в исторической науке[2].

Сегодня и в зарубежной литературе все чаще встречается точка зрения, что, например, «цифровая история» (digital history) имеет

Cohen D., Rosenzweig R. Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web. University of Pennsylvania Press, 2002.

[1] History in the digital age. London; New York: Routledge, 2013. В книге также делается занятное различение между «цифровыми историками» (специалистами, целенаправленно изучающими и внедряющими информационно-технологические решения в исторические исследования) и «историками в цифровую эпоху» (всеми профессиональными исследователями прошлого).

[2] Владимиров В. Н. Интернет для историка: и все-таки новая парадигма! // Круг идей: историческая информатика в информационном обществе: Труды VII конференции АИК. М., 2001; Володин А. Ю. «Цифровая история»: ремесло историка в цифровую эпоху // ЭНОЖ «История». 2015. Т. 6. № 8; Гарскова И. М. Историческая информатика. Эволюция междисциплинарного направления. СПб.: Алетейя, 2018.

важные отличительные черты, которые нежелательно смешивать с общими для гуманитарных наук трендами «дигитализации». Ярко эту мысль выразил С. Робертсон (директор Центра истории и новых медиа имени Р. Розенцвейга): рассматривая объединенные общей методологической платформой «цифровые гуманитарные науки» нельзя не заметить, что «источники, исследовательские вопросы и подходы, которые они используют в своих проектах, дисциплинарны, равно как дисциплинами определяется выбор цифровых инструментов»[1]. Такие мысли вполне созвучны российским дискуссиям на эту тему[2]. Например, М. Таллер разделил сообщество DH на четыре условные группы: 1) исследователи «текста как такового», 2) исследователи-собиратели «фактов» в электронных (иногда весьма обширных) коллекциях, 3) исследователи «нетекстов» (в том числе виртуальных реконструкций), 4) исследователи влияния цифровой среды на гуманитарные науки в целом[3].

В 2004 году был опубликован доклад о будущем historical information science (условно можно перевести как «об исторической информатике»), в котором основная перспектива развития виделась в сотрудничестве историков с учреждениями — хранителями историко-культурного наследия[4], так как именно в рамках такого сотрудничества могут сложиться основы для оцифровки и последующего компьютерного исследования исторических источников в самом широком смысле этого слова[5]. В тот момент историки лишь подходили к поиску «цифровой перспективы» исследований, о которой уже громко заявляли в работах филологи. Важной отправной точкой для дискуссии о развитии исследований в русле цифровых исследований в этой области стало издание «Компаньона по цифровым гуманитарным наукам», в котором были собраны многочисленные статьи, посвященные вопросам междисциплинарного синтеза в истории, филологии, археологии, антропологии и социальных науках. Фактически именно с выходом «Компаньона» в 2004 году

[1] Робертсон С. Различия между цифровыми гуманитарными науками и цифровой историей // Электронный научно-образовательный журнал «История». 2016. Т. 7. Вып. 7 (51). DOI: 10.18254/S0001648–1–1.

[2] Бородкин Л. И. Digital history: применение цифровых медиа в сохранении историко-культурного наследия? // Историческая информатика. 2012. № 1.

[3] Таллер М. Дискуссии вокруг Digital Humanities // Ист. информатика. 2012. № 1.

[4] Boonstra O., Breure L., Doorn P. Past, present and future of historical information science (Glasgow meeting, 25.04.2004). Amsterdam, 2004.

[5] Подробнее см.: McCrank L. J. Historical Information Science: An Emerging Unidiscipline. Information Today, 2002.

«цифровые гуманитарные науки» начали свое шествие по планете[1]. В 2016 году увидел свете уже «Новый компаньон по цифровым гуманитарным наукам»[2]. Оба компаньона вышли под редакцией трех исследователей — Сьюзан Шрейбман, Рея Сименса и Джона Ансворса. Еще в 2004 году редакторы называли «Компаньон» «поворотным пунктом в области цифровых гуманитарных наук», потому что «впервые широкий круг теоретиков и практиков, тех, кто работал в этой области в течение многих десятилетий, и тех, кто присоединился недавно, эксперты в разных дисциплинах, ученые-компьютерщики, специалисты в библиотечном деле и информационных исследованиях были объединены, чтобы рассмотреть цифровые гуманитарные науки как самостоятельную область знаний, а также задуматься о том, как она соотносится с традиционными гуманитарными исследованиями». Годы спустя редакторы замечают, что, хотя «остается спорным, следует ли рассматривать цифровые гуманитарные науки в качестве самостоятельной области знаний, а не всего лишь набора взаимосвязанных методов, но, без сомнения, в 2015 году цифровые гуманитарные исследования являются динамичной и быстро развивающейся областью научной деятельности» (p. xvii).

Редакторы «Нового компаньона» вспоминают, что сознательно решили отказаться от термина humanities computing и начали использовать название digital humanities с целью перенести ударение с компьютинга на гуманитарные науки. «Может быть, через десять или двадцать лет определение "цифровой" будет казаться излишним применительно к гуманитарным наукам. Возможно, по мере того, как все большая доля нашего культурного наследия будет оцифрована или уже создана в цифре (born digital), станет ничем не примечательным тот факт, что цифровые методы используются для изучения человеческого творчества, а мы будем думать об исследованиях, описанных в этой книге, просто как о "гуманитарных". Между тем редакторы этого "Нового компаньона по цифровым гуманитарным наукам" рады представить тщательно обновленный отчет о предметной области, как она существует сегодня».

«Новый компаньон» включает пять частей: инфраструктуры, создание, анализ, распространение и прошлое, настоящее, будущее

[1] A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, 2004. URL: http://www.digitalhumanities.org/companion/

[2] A New Companion to Digital Humanities / Eds Susan Schreibman, Ray Siemens, John Unsworth. Wiley-Blackwell, 2016. 592 p. ISBN: 978–1–118–68059–9.

цифровых гуманитарных наук. (Для сравнения: в компаньоне 2004 года было четыре части: история, принципы, приложения и производство, распространение, архивирование.) В разделе «Инфраструктуры» обсуждаются такие вопросы, как Интернет вещей, принципы коллективного использования данных и средств хранения оцифрованного культурного наследия. Раздел «Создание» посвящен особенностям междисциплинарных связей в цифровом контексте, новым медиа и вопросам моделирования, конструированию виртуальных миров и электронных библиотек. Раздел «Анализ» освещает моделирование данных, картографирование наблюдений, использование графических и мультимедийных форматов в цифровых исследованиях, анализ текстов и семантическую разметку. Раздел «Распространение» включает дискуссии о возможностях и ограничениях интерфейсов в цифровых проектах, о перспективах использования краудсорсинга и надеждах на разработку профессионального программного обеспечения для нужд цифровых гуманитарных наук. Раздел о положении в области описывает существующий научный ландшафт, показывает влияние глобализации и интернетофикации, обращает внимание на характерные черты цифровых исследовательских практик, прогнозирует ближайшие перемены в цифровой науке.

«Новый компаньон» показывает, что цифровые гуманитарные исследования переходят от теоретического самоопределения к научной практике — академическим открытиям и новым интерпретациям, памятуя об опасности оказаться «во власти программного обеспечения». Как справедливо замечает К. Уорвик, не стоит буквально понимать пословицу «больше вкалывай — меньше болтай» (more hack less yack) в отношении таких создающихся областей, как цифровые гуманитарные науки (p. 538). У. Томас поддерживает эту идею следующим размышлением: «Историки, литературные критики, философы, филологи, ученые, открывшие для себя цифровые гуманитарные исследования, начинают перестройку научной деятельности и ее организационных форм для нового цифрового мира. Ученые стали открытыми для самых различных исследовательских методик, для обмена источниками и материалами (данными), и признали крупномасштабные распределенные модели научных проектов. Ученые пришли к важному признанию, что мы сейчас живем в эпоху огромной емкости, вездесущего хранения, связанной сетевой информации и беспрецедентного доступа. Вместо привычной манеры исследований, ориентированных на редкие материалы,

к которым ограничен доступ, а эксперты самостоятельно проводят его отбор, цифровые гуманитарные науки в своих наиболее ярких проявлениях основываются на расширении сферы гуманитарных исследований, открывая доступ к источникам, а также обогащая понятие научной деятельности» (p. 524).

Любопытно, что руководства и дискуссионные тома, вышедшие вслед за «Компаньонами» уже фокусируются не на подходах и инструментах исследования, а на критическом рассмотрении контекстов, в которых эти исследования ведутся, погружая дискуссии о цифровых гуманитарных исследованиях в широкую философскую и политическую парадигму[1].

Цифровые методы получают все большее распространение в сфере гуманитарных наук и приобретают свои организационные формы в виде специализированных конференций, исследовательских центров и журналов.

Свои исследовательские центры, действующие в области цифровых гуманитарных наук, есть у множества крупных университетов по всему миру. Статус наиболее авторитетного в сфере цифрового литературоведения удалось получить Стэнфордской литературной лаборатории. В исторических исследованиях широкую известность приобрели Центр истории и новых медиа имени Роя Розенцвейга (RRCHNM), Люксембургский центр современной и цифровой истории (C2DH).

В России соответствующие институции имеются в Москве (МГУ, НИУ ВШЭ), Санкт-Петербурге (ИТМО, Институт русской литературы РАН), Барнауле (АлтГУ), Екатеринбурге (УрФУ), Калининграде (БФУ), Красноярске (СФУ), Перми (НИУ ВШЭ-Пермь), Ростове-на-Дону (ЮФУ), Томске (ТГУ).

К зарекомендовавшим себя авторитетам научной периодики относятся *Digital scholarship in the humanities* (с 2012 года, с 1986 года выходил под названием Literary and Linguistic Computing), *Digital humanities quarterly* (с 2007 года), Историческая информатика (с 2012), отчасти *Journal of Cultural analytics* присоединяются *International Journal of Digital Humanities* (с 2019 года), Квантитативная филология (с 2021 года), *Journal of Digital History* (с 2021 года), *Digital Orientalia* (с 2021 года), *Journal of Computational Literary Studies* (с 2023 года), Цифровые гуманитарные исследования (с 2023 года).

[1] The Bloomsbury Handbook to the Digital Humanities / Ed. James O'Sullivan. Bloomsbury Academic, 2022. 512 p.; Debates in the Digital Humanities 2023 / Eds. Matthew K. Gold and Lauren F. Klein. Univ Of Minnesota Press, 2023. 520 p.

Важными для определения границ научного поля становятся конференции, центральное место среди которых занимает ежегодная созываемая ассоциацией организаций цифровых гуманитарных наук (ADHO).

Состав секций и круглых столов этой конференции лучше всего описывает ключевые направления, в которых работают представители цифровых гуманитарных наук:

- самоопределение digital humanities;
- проблемы оцифровки, организации электронных изданий и публикации данных;
- методы, в частности, обработка естественного языка, сетевой анализ, компьютерное зрение.

С исследовательской точки зрения DH — это проектный подход к решению научных проблем, предполагающий в качестве итога исследовательского труда конкретный информационный цифровой продукт (набор данных, онлайн-ресурс, информационную систему). Проектный подход в том числе означает и соавторство, участие в исследовании нескольких авторов, каждый из которых вносит свой вклад, соразмерный его компетенциям. Действительно, обычная для цифровых гуманитарных наук картина — это наличие нескольких авторов у одной публикации (пример этому — данная монография). В то же время известно, что у абсолютного большинства статей на гуманитарную тематику только один автор[1].

С образовательной точки зрения цифровые гуманитарные науки можно рассматривать как привлекательное для студентов направление обучения. В таком смысле DH — это комплекс дисциплин, позволяющих представить специфику изучения гуманитарных проблем в современных условиях, то есть в эпоху «больших данных» и исследовательских облачных платформ. Как показывает практика, образовательные программы DH весьма популярны (как на бакалаврском, так и на магистерском уровне) в США, Великобритании, Германии, Франции, Японии, Австралии. При этом дисциплины, преподаваемые в соответствующих циклах, имеют существенный технологический уклон, не теряя очевидной возможности включения гуманитарных знаний и исследований в актуальную мультимедийную среду. Можно сказать, что уже складывается определенный образовательный DH-канон — оцифровка, модели и базы данных,

[1] Жэнгра И. Ошибки в оценке науки, или Как правильно использовать библиометрию. М.: НЛО, 2018. С. 44.

метаданные и разметка, интеллектуальный и сетевой анализ, визуализация и картографирование данных, трехмерное моделирование, веб-ресурсы и интерфейсы, проектный подход и интеллектуальная собственность[1]. При этом все перечисленные знания и умения не отменяют необходимости глубоко разбираться в конкретном предмете гуманитарного исследования.

И, наконец, с точки зрения профессионального сообщества, DH — это полезный бренд, позволяющий обращаться за финансированием и административной поддержкой, предлагая инновационные решения для вполне классических гуманитарных дисциплин. В последнее время, особенно в контексте цифровизации и успехов алгоритмических решений в повседневной жизни, наблюдается значительный общественный интерес к результатам реализации проектов в области DH, в том числе и потому, что присутствие такого рода проектов в интернете делает их более доступными любопытствующей публике. Вместе с тем онлайн-публикация результатов исследований способствует и международным дискуссиям. Такого рода перемены свидетельствуют о прямой практической пользе от «цифрового поворота».

Не будем скрывать скептического отношения многих традиционных гуманитариев, подмечающих и методологические слабости молодой компьютерной области, и легковесность стоящих за цифровыми исследованиями концепций, и неоднозначность продуцируемых выводов. Однако междисциплинарное направление DH за два десятилетия не только заявило о себе, но и утвердилось, нашло свою нишу в исследовательском сообществе и сформировало подход к реализации гуманитарных исследовательских проектов в цифровую эпоху. Необходимо подчеркнуть, что цифровые гуманитарные науки сформулировали цели и задачи в конкретный исторический момент — значительного увеличения вычислительных возможностей и расширения сетевого международного взаимодействия, при этом не превратились в закрытый клуб исследователей. По сути, современные цифровые гуманитарные науки предполагают широкую исследовательскую программу, которая включает вопросы, интересующие любого гуманитария. Цифровые исследовательские практики — это реальность любого ученого.

[1] См., например: Drucker J. The Digital Humanities Coursebook. An Introduction to Digital Methods for Research and Scholarship. Routledge, 2021.

Предсказывать вектор развития DH сложно. Эта область не вполне самостоятельна: ее будущее зависит от развития технологий, и траектории будущего дисциплины находятся в прямой связи с появлением и проработкой новых методов. Еще в начале 2010-х годов было крайне трудно распознать, что именно искусственные нейронные сети станут настолько эффективным инструментом решения множества интеллектуальных задач. Нет возможности предсказать, какие именно технологии появятся в ближайшем будущем. Но все же в общих чертах ясно, что главный курс дальнейшего развития DH связан с применением различных методов анализа гуманитарных данных и расширением разнообразия результатов таких исследований. Цифровые гуманитарные науки будут двигаться в сторону моделирования все более сложных и трудноформализуемых объектов и уровней. Значительный шаг в этом направлении — моделирование семантики, реализованное в форме векторных моделей (см. главу об анализе текста).

Глава 2

Данные в цифровых гуманитарных исследованиях

(*А. Ю. Володин*)

Настоящее всегда чревато будущим.
Готфрид Вильгельм Лейбниц

В 1703 году Г. Лейбниц описал двоичную систему счисления с 0 и 1. Благодаря непосредственной реализации в цифровых электронных схемах на логических вентилях двоичная система используется практически на всех современных компьютерах. В научных книгах современный этап развития гуманитарных знаний часто называют «цифровым поворотом» — период, когда тексты, изображения, звуки стали переходить из аналогового формата в цифровой, пришелся на рубеж XX–XXI веков[1]. С формальной точки зрения цифровой формат — это тип сигналов данных в электронике, использующих дискретные состояния (в отличие от аналогового сигнала, использующего непрерывные изменения сигнала). С содержательной точки зрения цифровой формат — это новый способ создания, преобразования, накопления, передачи и использования информации как в науке, так и в быту. В настоящий момент мы живем в период «перекодировки», назовем условно, аналоговых документов в цифровые[2]. Такую «перекодировку» часто называют «оцифровкой». Оцифровка, по определению, является описанием

[1] Viola Lorella. The Humanities in the Digital: Beyond Critical Digital Humanities. Palgrave Macmillan, 2023.

[2] Подробное исследование проблемы «перекодировки» см.: Switching Codes: Thinking Through Digital Technology in the Humanities and the Arts / T. Bartscherer, R. Coover (eds.). University of Chicago Press, 2011.

объекта, изображения или аудио-видеосигнала в виде набора дискретных цифровых замеров этого объекта/сигнала при помощи специальной аппаратуры — иными словами, перевод в цифровой вид, пригодный для записи на электронные носители.

Очевидно, что перед исследователем встает вопрос: зачем оцифровка нужна? Ответ кажется очевидным: оцифровка позволяет получить, благодаря развитию сетевых технологий обмена информацией, быстрый доступ к цифровым комплексам информации, когда доступ к аналоговым комплексам потребовал бы существенно больше времени и средств. При дальнейшем же размышлении важно ответить на важный вопрос: а есть ли у оцифровки еще какие-то преимущества для исследователя, кроме быстроты доступа к оцифрованным материалам, и даже шире — что оцифровка исследователю дает и что при оцифровке исследователь теряет?

Важным следствием набирающей скорость оцифровки стало появление в открытом доступе широкого комплекса текстов и изображений, доступных для поиска, анализа и обобщения. Доступность, которую нам дарят электронные ресурсы, не является синонимом легкости их использования исследователем, потому что удобный доступ не снимает вопросов критического отношения к источнику, тем более, если это электронная копия, аутентичность которой еще надо установить. Историки включились в процесс «цифрового перехода» в конце XX в., когда компьютеризация перестала быть связана только с квантификацией, но и распространилась на историческую эвристику, критику и интерпретацию.

Изменение средств и форматов передачи информации существенно повлияло на практику историков. Оцифровка сегодня стала привычной практикой любого историка, будь она спонтанная, когда нужно «выписать» важную цитату, или систематическая, когда оцифровывается комплекс документов. Оцифровка, очевидно, усложнилась: помимо получения цифрового образа документа, необходимо распознавание образа, кодирование информации, а затем и формирование целостного информационного комплекса с метаинформацией. Например, Я. Грегори утверждает, что современное поколение историков — первое, которое сталкивается с сложными и неполными цифровыми (или оцифрованными) источниками, кишащими ошибками оцифровки, распознавания и кодирования[1].

[1] Gregory Ian. Challenges and opportunities for digital history. Front. Digit. Humanit. 1:1. 2014. DOI: 10.3389/fdigh.2014.00001.

Многие исследования начинаются с оцифровки. По сути, оцифровка для исследователя становится опытом коллекционирования, формирования собственной базы знаний, позволяющей нетривиально связывать наблюдения и гипотезы. Опасения, связанные с оцифровкой, основаны на справедливом предположении, что современные технологии оцифровки (в частности, сканирования) не позволяют сохранить в цифровом формате полноценную виртуальную копию каждого объекта прошлого.

Два режима — аналоговый и цифровой — сосуществуют, при этом есть пути перевода из одного в другой (например, оцифровка или распечатка). Каждый информационный ресурс имеет свои черты по природе происхождения — именно по этой причине так часто высказывается недоверие к результатам оцифровки бумажных документов. Очевидно, что оцифровка — это определенное абстрагирование от оригинала. Но это абстрагирование при верном научном подходе может принести немало пользы. Неверно обращать внимание исключительно на быстроту доставки файла оцифрованного исторического источника. Важно заметить, что оцифровка — это кодирование, а кодирование вместе с систематическим описанием может стать прочной основой для исторической реконструкции.

Цифровые сигналы существуют как последовательности чисел во времени. Часто для цифрового кодирования достаточно двух чисел — 0 и 1, которые называют битами. В таком смысле «цифровой» способ хранения данных обозначает цифровой (двоичный) формат. Однако на практике важно обращать внимание на различия цифровых форматов. Ведь каждый цифровой формат в состоянии хранить (как контейнер) разное количество информации. Получается, что цифровые форматы позволяют создавать многослойную информацию, когда один файл может содержать и текст первоисточника, и варианты его критики и интерпретации (как слои в файлах изображений).

И возникает не менее важный вопрос: является ли оцифровка практикой исторических исследований или же это удел специалистов по построению информационных систем?[1] Нельзя сказать, что есть простой ответ. Кратко обобщая мнения, встречающиеся в литературе, отметим, что оцифровка может оказаться полезной

[1] См., например: Яник А. А. Исторические исследования в новых реальностях информационного общества XXI века // История современной России: Цифровая инфраструктура междисциплинарных исследований. М.: Изд-во Моск. ун-та, 2014. С. 11–38.

и осмысленной для целей исторической науки, если она целенаправленно решает три задачи:

– сохраняет важные исторические документы в цифровых форматах[1];

– позволяет применить электронные средства обработки данных (как ручные, так и автоматизированные)[2];

– представляет документальную основу для исследований заинтересованной общественности, например, онлайн[3].

Для плодотворного развития «цифровой истории» важно рассматривать оцифровку как повседневную исследовательскую практику, постепенно гарантирующую переход от спонтанной к систематической оцифровке, а также переход к расширительному пониманию источников исторического исследования — их полному цифровому представлению, изучению не только текстов, но и памятников прошлого в их цветущей мультимедийной сложности. Такой вариант развития означает переход от оцифровки как описания (записи в новом формате, реплики, копии) к оцифровке как реконструкции (созданию новых контекстов, порождению новых смыслов). Вместе с этим момент требует установления стабильных инструментов и методик, позволяющих использовать сильные стороны цифровых форматов представления данных, чтобы «цифровой переход» оказался не только повторением или продолжением классических практик новыми техническими средствами, но и позволил бы открыть новые горизонты познания и понимания.

Таким образом, оцифровку можно рассматривать как кодирование. Кодирование — это преобразование сигнала из формы, удобной для непосредственного использования информации, в форму, удобную для передачи, хранения или автоматической обработки. Именно по этой причине оцифровка оказалась в центре дискуссий о судьбе «цифровой истории». Кодирование, что важно, может служить не только целям автоматизации обработки информации,

---

[1] См., например: Бородкин Л. Digital history: применение цифровых медиа в сохранении историко-культурного наследия? // Историческая информатика. 2012. Т. 1. № 1. С. 14–21.

[2] В данном контексте можно отметить дискуссию вокруг книги Франко Моретти «Дальнее чтение» (М.: Изд-во Института Гайдара, 2016), в которой практика «близкого чтения» (close reading) противопоставляется «дальнему чтению» (distant reading), состоящему в количественном и систематическом анализе большого корпуса текстов (комплекса документов).

[3] Важный уникальный отечественный пример — корпус дневников и воспоминаний «Прожито»: https://prozhito.org/

но и становиться системой описания. Описание же –одно из важных средств познания в истории прежде всего потому, что описание в историческом исследовании может стать одним из путей реконструкции прошлого — его контекстуализации, операционализации, связывания «больших данных».

Эра данных

> Датаизм провозглашает, что Вселенная состоит из потоков данных и что ценность всякого явления или сущности определяется их вкладом в обработку данных.
>
> *Юваль Ной Харари. Homo Deus*

Данные — это подход к регистрированию явлений действительности, претендующий на абсолютную формализацию не только хранения зарегистрированной информации, но и ее познания. Регистрируемость явлений и событий — сложная проблема в гуманитарном познании.

Многообразие существующих и создающихся электронных ресурсов переносит нас в «эру данных». Данные меняют подход к исследовательским материалам хотя бы потому, что они оказываются недоступными человеку без специального устройства-посредника (недаром данные часто и сегодня называют машиночитаемыми)[1]. Такого рода перемены вносят существенные изменения и в исследовательские практики. Правда, влияние новых средств коммуникации на информационную среду замечено было давно. Еще М. Маклюэн выделял период развития медиасреды — «галактику Маркони», которая пришла на смену «галактике Гутенберга» уже больше века назад, с приходом электричества в повседневную коммуникацию[2].

Датафикация — процесс устойчивого фиксирования массовых наблюдений в разных форматах данных, позволяющий осуществлять

[1] Маккарти У. Специальные эффекты: инструменты есть, а где результаты? // Электронный научно-образовательный журнал «История». 2016. Т. 7. Вып. 7 (51). URL: https://history.jes.su/s207987840001637–9–1/ (дата обращения: 26.08.2023).

[2] Маклюэн М. Галактика Гутенберга. Становление человека печатающего. М.: Б. и., 2005. 496 с.

их качественную и количественную обработку и научный анализ. Измерение (то есть установление соотношения качественных и количественных характеристик) объектов, явлений, процессов реального мира и запись получаемых данных — важная характеристика практически всех обществ письменной истории.

По сути, датафикация — общий для науки процесс, который может протекать одинаково в разных гуманитарных дисциплинах. Хотя каждый специалист будет настаивать на принципиальных отличиях данных в собственной научной области. Как например С. Робертсон отмечает, что несмотря на общую методологическую платформу «цифровые гуманитарные науки», «источники, исследовательские вопросы и подходы, которые они используют в своих проектах, дисциплинарны, равно как дисциплинами определяется выбор цифровых инструментов»[1].

Компьютерная датафикация имеет длительную и насыщенную традицию. «Появление компьютеров повлекло за собой внедрение цифровых устройств для измерения и хранения данных, которые значительно повысили эффективность датафикации, — пишут В. Майер-Шенбергер и К. Кукьер, — а также сделали возможным математический анализ данных для раскрытия их скрытой ценности. Проще говоря, оцифровка стала катализатором датафикации, но никак не ее заменой. Процесс оцифровки (преобразование аналоговой информации в формат, считываемый компьютером) сам по себе не является датафикацией»[2].

С научной точки зрения датафикация — это процесс нормализации наблюдений для их систематического анализа. Причем с учетом того, что современные подходы позволяют работать как со структурированными, так и со слабо структурированными или вовсе не структурированными данными, уместно говорить не о структурировании данных в соответствии с «нормальной формой», а о гармонизации данных для решения конкретных исследовательских задач. Гармонизация данных предполагает проведение комплекса мероприятий по повышению степени их согласованности. Сначала процесс гармонизации осуществляется на семантическом уровне,

[1] Робертсон С. Различия между цифровыми гуманитарными науками и цифровой историей // Электронный научно-образовательный журнал «История». 2016. Т. 7. Вып. 7 (51). URL: https://history.jes.su/s207987840001648–1–1/ (дата обращения: 26.08.2023).

[2] Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. М.: Б. и., 2014. С. 83.

а затем анализируются технологические возможности и ограничения форматов хранения данных в файловой структуре.

Получается, что датафикация оказывается успешной в том случае, когда полученные из объектов исследования данные оказываются удобоваримыми для автоматизированного компьютеризированного использования, анализа и управления.

Что такое данные и кем они даны?

> 4.259 **данные**: предоставление информации в формальном виде, пригодном для передачи, интерпретации или обработки людьми или компьютерами.
>
> *(ISO/IEC2382:2015)*

Сегодня часто можно услышать, что мы живем в мире данных, что данные — это новая нефть, что нас окружают «большие данные». Давайте начнем с простого: ответим на вопрос, что такое данные? Данные, если обратиться к стандартному определению, это предоставление информации в формализованном виде, пригодном для передачи, интерпретации или обработки людьми или компьютерами. Также можно встретить определение, что данные — это интерпретируемое представление информации в форме, удобной для передачи, интерпретации и обработки.

Здесь важно отметить, что речь идет о формальном представлении информации об окружающем нас мире. И форма имеет значение, потому что в зависимости от того, как именно представлены данные, зависит то, что мы сможем с ними сделать.

Главное, что важно уяснить с методической точки зрения: данные — это способ абстрагирования, то есть способ отвлечься от многообразия свойств изучаемого объекта и зафиксировать ряд свойств, которые кажутся принципиальными, важными для наблюдающего объект исследователя.

Данные — это абстракции сущностей реального мира (человека, объекта или события). Данными могут стать любые зафиксированные сигналы. Данные состоят из свойств, или переменных, или атрибутов, как бы мы ни привыкли называть конкретный вид абстракции. Каждый объект обычно описывается рядом атрибутов.

Простой пример, хорошо известный любому гуманитарию, — книга, которая может иметь следующие свойства: автор, название, издательство, место и год издания, количество страниц, ISBN, цена, жанр, тема и т.д. Важным навыком цифрового гуманитария является умение рассматривать текст как данные[1].

Обычно мы имеем дело не со случайными данными, а с наборами данных, специально подготовленными для решения конкретной задачи: это может быть классификация, может быть интерпретация, может быть выявление тождества, закономерности или противоречий.

Для простоты обратимся к примеру: на всех цифровых устройствах от компьютера до смартфона у нас хранятся файлы. Файл — это поименованная совокупность записей, рассматриваемая как единое целое. Или, если сказать более технично, файл — именованная область данных на носителе информации. Мы уже автоматически различаем разные типы файлов — этот файл текстовый, а этот — изображение, это — видеофайл, а этот — аудиофайл с любимой музыкальной композицией.

Условно можно выделить несколько типов цифрового представления информации: текст, таблица и изображение, затем динамические потоковые форматы: аудио и видео, а далее — программные коды. Особенность «цифрового поворота» можно увидеть в том, что указанные типы на практике чаще всего создают неожиданные сочетания, к примеру: оцифрованная средневековая рукопись первоначально предстает в виде цифрового изображения, затем проводится процедура установления текста, даже если в ней и принимают участие компьютерные технологии, например, таблицы базы данных, что в основном согласуется с классическим текстологическим подходом. Таблица придает структуру информации самого разного характера, а связанные таблицы уже могут превратиться в базу данных. Поэтому важным свойством, которое позволяет увидеть перспективы цифрового подхода, является многослойность цифровых форматов. Необходимость сочетать структурированную и графическую информацию, слабоструктурированную или неструктурированную информацию дает возможность создать основу для мультимедийного

[1] См., например: Grimmer Justin, Roberts Margaret E., Stewart Brandon M. Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press, 2022. 360 p.; Blaney Jonathan, Winters Jane, Milligan Sarah, Steer Martin. Doing digital history: A beginner's guide to working with text as data. Manchester University Press, 2021. 192 p.

понимания современных электронных ресурсов. Многообразие форматов современной передачи информации во многом скрывает от взгляда существенные различия их информационного потенциала, необходимые для актуального осмысления, например, то, что один и тот же файл может сообщить разное количество информации в зависимости от программы, в которой он открыт. Получается, что исследователь оказывается в прямой зависимости от функционала программного обеспечения. Однако в основе результата процесса оцифровки — именно данные.

Сегодня есть сотни проектов, посвященных оцифровке историко-культурного наследия, и работать с такими наборами данных, коллекциями, репозиториями нужно уметь. Эти данные доступны, но к ним нужно уметь подойти, обратиться и увидеть те закономерности, которые позволяют открыть современные методы цифровых гуманитарных исследований. Именно это является одной их ключевых задач в преподавании цифровых гуманитарных наук.

В науке сложилось представление о так называемой информационной иерархии: в основе ее располагается сигнал, формализация сигналов создает данные, из данных мы получаем информацию, информация складывается в наши знания, а знания, наконец, позволяют нам понять окружающий мир. Данные оказываются в основе порождения смыслов. Данные, которым мы добавляем содержательный контекст, становятся информацией. Когда мы знаем, как поступить с информацией, мы обладаем знанием, а когда мы учитываем возможности и границы применимости знания, мы уже переходим на уровень понимания.

Сегодня стала широко известной наука о данных, изучение данных с целью извлечения значимой информации, междисциплинарный подход, который сочетает в себе принципы и методы математики, статистики, машинного обучения для анализа больших объемов данных. При этом стоит заметить, что Комитет по данным Международного научного совета (или CODATA) был создан еще в 1966 году. Один из пионеров информатики Петер Наур определял компьютерные науки как даталогию, которая изучает жизненный цикл научных данных — от появления до преобразования.

Почему данные бывают большими и малыми, какие из них лучше?

> В любом наборе исходных данных самая надежная величина, не требующая никакой проверки, является ошибочной.
>
> *Третий закон Финэйгла*

Итак, данные — форма представления зафиксированных сигналов, пригодная для обработки пользователями и информационными системами. Смысл такой формализации в том, чтобы решить задачи передачи, обработки, интерпретации зафиксированных сигналов.

Когда говорят про эпоху данных, могут иметь в виду разные значения данных. Во-первых, данные — это **формат**, то есть представление фактов и идей в формализованном виде, пригодном для передачи и обработки в некотором информационном процессе. Во-вторых, данные — это **память**, то есть совокупность ячеек памяти, обладающих определенными свойствами. И наконец, в-третьих, данные, в отличие от операций (действий или процессов), **номинальны**, и, чтобы получить какой-то результат, как раз необходимо провести какое-то исследовательское действие с ними.

Нельзя не отметить, что в последние годы в науке и общественных дискуссиях все чаще встречается понятие «большие данные». Всплеск интереса к таким данным возник на волне цифровизации, когда буквально каждый человек стал регистратором многочисленных процессов — от геолокации до общественного мнения.

Стоит иметь в виду, что под большими данными понимают не столько сами данные, сколько подходы, инструменты и методы обработки больших объемов данных. К большим данным относят как структурированные, так и неструктурированные, и, что особенно важно и сложно, неопределенно структурированные данные огромных объемов и значительного многообразия. Главная задача подхода к большим данным — получить воспринимаемые человеком результаты, учитывая непрерывный прирост таких данных, распределенных по многочисленным узлам вычислительных сетей.

Но говоря о гуманитарных данных, надо иметь в виду, что они далеко не всегда большие. И это ни плохо, ни хорошо. Данных должно быть достаточно, чтобы с их помощью можно было бы решить ту или иную исследовательскую задачу. В книги Кристины

Боргман как раз обсуждается соотношение больших данных с данными малыми. Но еще важнее, стоит признать, что в гуманитарных исследованиях часто складывается ситуация, что данных просто нет и добыть их неоткуда. Отсутствие данных также важно учитывать в современной методологии.

Исследование данных сегодня часто называют интеллектуальным анализом данных. Суть данного подхода — в поиске нетривиальных связей в данных. К данным применяются различные методики классификации, моделирования, прогнозирования, помогающие увидеть в данных смысл. Данные сами не говорят, но тот, кто умеет их расспросить, сможет многое узнать.

Уместно вспомнить замечательное рассуждение Ю. М. Лотмана о значении дешифровки для исторической профессии: «Историк обречен иметь дело с текстами. Между событием “как оно есть” и историком стоит текст, и это коренным образом меняет научную ситуацию. Текст всегда кем-то создан и представляет собой происшедшее событие, переведенное на какой-то язык. Одна и та же реальность, кодированная разными способами, даст различные — иногда противоположные — тексты. Извлечение из текста факта, из рассказа о событии — события представляет собой операцию *дешифровки* (выделено мной. — *А. В.*). Таким образом, сознавая это или нет, историк начинает с семиотических манипуляций со своим исходным материалом — текстом»[1]. При этом сегодня мы смотрим на «текст» как на существующий в нескольких состояниях объект — рукопись, установленный печатный вариант рукописи и распознанный посимвольно электронный текст в цифровом файле инструментально дают разные возможности для исследователя.

[1] Лотман Ю. М. Изъявление Господне или азартная игра? (Закономерное и случайное в историческом процессе) // Ю. М. Лотман и тартуско-московская семиотическая школа. М., 1994. С. 353–354.

Что такое ка́пта и почему гуманитарии часто работают только с тем, что собрали сами?

Процесс научных открытий — это, в сущности, непрерывное бегство от чудес.

Альберт Эйнштейн

Данные — даже по названию — создают впечатление, что они кем-то даны. В действительности сбор данных требует много труда. Данные (особенно гуманитарные) появляются благодаря скрупулезному повседневному труду. «Эра данных» в гуманитарных исследованиях и смежных гуманитарных дисциплинах синхронна общенаучным тенденциям обсуждения «больших данных» (big data). Большие массивы данных требуют новых подходов, при этом специализация технологических решений для нужд гуманитарного исследования принципиально необходима. Однако общий подход к большим данным часто вступает в противоречие с определением больших данных как потоковых и постоянно пополняемых массивов, а значит, в историческом или гуманитарном исследовании целесообразно использовать прежде всего понятие средних (medium) или даже малых (small) данных, ведь для задач исследования важен не объем данных как таковой, а их достаточность для обоснованного вывода. В современной историографии все чаще возникают дискуссии о роли данных в актуальных исторических исследованиях, которые так или иначе превращаются в обсуждение роли «больших данных» в современной науке, в том числе в истории[1].

В качестве определяющих характеристик для «больших данных» обычно принято выделять пять свойств (т.н. «пять V»): объем (volume) в смысле величины физического объема для хранения данных, скорость (velocity) прироста, а значит скорость обработки и получения результатов, многообразие (variety) как возможность одновременной обработки различных типов структурированных и полуструктурированных данных. В последнее время к классической тройке формальных свойств присоединились и важные содержательные характеристики: достоверность (veracity) и изменчивость

[1] Guldi J., Armitage D. The History Manifesto. Cambridge University Press, 2014. 175 p.

(variability). Как показывает опыт последних лет, «большие данные» требуют нескольких этапов сотрудничества различных специалистов: 1) получение данных (в том числе с помощью популярного сетевого краудсорсинга, под которым часто понимают привлечение множества волонтеров к трудозатратной деятельности, которую можно выполнить онлайн), 2) документирование данных (в том числе мета-описания, курирования и гармонизации), 3) обработка данных (включая агрегирование и «добычу» данных), 4) анализ данных (позволяющий создавать модели и теории), 5) визуализация данных (в частности, создания интерфейса для запросов и выдачи результатов обработки данных).

«Средние данные» (medium data) все чаще применяются в научной литературе для описания крупных коллекций данных, которые не претендуют на поток постоянных пополнений в духе «больших» данных. Тем не менее использование данных средних объемов может оказаться наиболее удобным при апробации новых подходов, так как они позволяют строго контролировать имеющиеся переменные и наблюдения и создают возможности для построения взаимосвязей — графиков, карт (например, на платформе Tableau или в среде RStudio), сетевого анализа (например, Ucinet и NetDraw), а также автоматической семантической разметки (например, с помощью MaxQDA или n-gram). Недостаток же «средних данных» состоит в том, что обычно они отбираются из больших коллекций по одному или нескольким техническим параметрам, и решение об их репрезентативности принимает исследователь.

«Малые данные» (small data) все чаще употребляются в качестве противопоставления: «малые данные», собранные исследователем самостоятельно (как раз речь идет о капте), отличаются от «больших» и «средних» данных, полученных (скачанных/загруженных) из различных депозитариев научной информации. Обращение к «малым данным» позволяет проявить исследовательские компетенции при построении инфологической и даталогической моделей. «Малые данные» полностью контролируются исследователем, при этом обычно подробно обосновываются репрезентативность и целостность самостоятельно собранных данных, а для нужд вторичного использования осуществляется подробное документирование исторических источников таких коллекций данных.

Споры о данных стали важным этапом восприятия историческим сообществом нового этапа развития информационных технологий, а с ними значительно возрос интерес к наработанным

подходам и признанным технологиям в рамках таких направлений, как историческая информатика.

В «Историческом манифесте» Д. Гулди и Д. Армитедж определяют причину «кризиса долгосрочного мышления», а вместе с ним и интереса к истории, информационной перегрузкой современных людей. В частности, они отмечают: «Мы живем в новую эпоху “больших данных” — от расшифровки генома человека до миллиардов слов в официальных отчетах, которые ежегодно производят правительственные учреждения. В социальных и гуманитарных науках обращение историков и социологов к “большим данным” отражает их стремление идти в ногу со временем, использовать открывающиеся возможности для решения старых вопросов и формулирования новых. “Большие данные” стимулируют социальные науки к постановке более масштабных проблем. В истории это, прежде всего, события мирового масштаба и длительная институциональная динамика. В проектах, посвященных долгосрочной истории изменений климата, последствиям работорговли или разнообразию форм права собственности на Западе, использование вычислительных методов позволяет исследователям открывать новые аспекты работы с данными и связывать исторические проблемы с современными»[1].

Стоит заметить, что данное поветрие в мировой науке стимулирует вовлечение в дискуссию самого широкого круга участников, часто весьма условно представляющих роль данных в целом и баз данных в частности, фактически уже многие десятилетия использующихся в исторической науке[2].

В гуманитарных науках все чаще начинает использоваться понятие «к***а***пта». Что такое капта? Образно говоря, это исследовательский «улов». Это те данные, которые историк собрал в архиве, лингвист, например, в поле, а философ на ментальной карте. Иногда такие данные называют «естественной выборкой», понимая под ней те оставшиеся источники, документы, артефакты прошлого, которые мы можем анализировать. Фактически это собранные доступные данные.

---

[1] Guldi J., Armitage D. The History Manifesto. Cambridge: Cambridge University Press, 2014. P. 88.

[2] Гарскова И.М. Историческая информатика. Эволюция междисциплинарного направления. СПб.: Алетейя, 2018. 408 с.; Бородкин Л.И. Моделирование исторических процессов: от реконструкции к анализу альтернатив. СПб.: Алетейя, 2016. 310 с.

Для понимания капты может помочь образ археологического раскопа. То, что найдено в раскопе в этом году, является последней по близости к настоящему находкой, но лишь очередной на поступательном пути науки. В следующем году будут новые находки, но анализировать и интерпретировать мы можем только то, что есть у нас в руках сегодня. Да, будущие открытия дополнят наши знания, но исследуем мы то, что есть.

Представьте, что вы набираете сообщение другу. Каждая буква отличается от другой, но алфавит давно вам знаком, и вы, не задумываясь (хотя физиологически мозг решает эту задачу, пусть почти автоматически), складываете буквы в слова, слова в предложения, предложения в абзацы. Вы формулируете мысли, информируете, побуждаете. При этом можете не задумываться о том, какая система кодирования нужна, чтобы передавать сообщения по каналам интернет-связи, да еще и так, чтобы они были получены так же, как и отправлены. Как можно не знать, что говоришь прозой, так же можно и не знать, что печатаешь благодаря стандарту кодирования символов письменных языков Unicode.

Так мы самостоятельно порождаем наборы данных. Да, сиюминутных, часто случайных, повседневных. Причем создаем их все больше и больше. И если в 1956 году жесткий диск на пять мегабайт был огромным, то сегодня мы часто недовольны качеством фотографии, если она «весит» всего пять мегабайт.

При оцифровке объектов историко-культурного наследия качество имеет первостепенное значение, и иногда цифровые копии объектов оказываются грандиозных размеров. К примеру, знаменитая цифровая копия «Ночного дозора» сделана высокоточной камерой и состоит из 717 миллиардов пикселей размером в 5,6 терабайта.

Но не только сами изображения являются данными, но и их описания. К примеру, известная электронная библиотека «Европеана» каждый хранящийся объект подробно описывает. Такое описание называется метаданными. Метаданные — это данные о данных. Например, все свойства картины или скульптуры, которые позволяют ее в деталях описать. Важно понимать, каким образом устроены те или иные данные и метаданные, чтобы иметь возможность ставить исследовательские задачи. По этой причине сегодня коллекции данных требуют не только береженого хранения, но и хранителей, которых часто называют кураторами данных.

В пример можно привести репозиторий открытых данных по русской литературе и фольклору Института русской литературы

РАН, где хранятся именно наборы данных, необходимые для проведения количественных литературоведческих исследований.

Создатель Всемирной паутины, автор концепции семантической сети и инициатор подхода «связанных данных» Тим Бернерс-Ли предложил так называемую 5-звездочную схему развертывания открытых данных. Самое простое — сделать свои материалы доступными онлайн в любом формате (за такие данные дается одна звезда). Если свои данные структурировать, например, в табличном виде, то это уже достойно оценки в две звезды. Перевод данных в свободные (непроприетарные) форматы (как, например, XML) — это уже следующий шаг, достойный трех звезд. Самые подготовленные данные, такие, которые используют унифицированный идентификатор (так называемый URI, позволяющий указывать на любой онлайн-ресурс: документ, изображение, файл), получают четыре звезды. И наконец, пяти звезд могут удостоиться связанные данные, или linked oped data (LOD), благодаря чему можно связывать друг с другом самые разные коллекции данных.

Для исследовательских задач работы с данными требуется внимательное отношение к семантическому моделированию данных[1]. Такой подход строится на понимании смысла этих данных. Причем различные исследователи могут вкладывать в одни и те же данные разные смыслы, видеть разные связи, проверять различные гипотезы.

Как мы уже говорили, гуманитарные данные часто называют каптой — исследовательским уловом, который ученый добыл своим трудом. Почему же сегодня столь популярной стала концепция FAIR по отношению к исследовательским данным? Акроним FAIR описывает так называемые честные данные. Такие данные отличают четыре свойства. В честных данных можно осуществлять поиск. Честные данные должны быть доступны. Такие данные должны быть интероперабельны, то есть совместимы с разными программами и операционными системами. И наконец, FAIR-данные должны поддаваться повторному или многоразовому использованию.

Почему это действительно важно? В идеальном мире каждое новое исследование должно привносить в копилку науки новые наборы данных, базы данных, оцифрованные документы и артефакты. И все эти материалы, как кусочки смальты, шаг за шагом должны создавать или воссоздавать огромное мозаичное панно истории

[1] Полезное руководство, позволяющее избежать ошибок при проектировании моделей данных: Alexopoulos Panos. Semantic Modeling for Data: Avoiding Pitfalls and Breaking Dilemmas. O'Reilly Media, 2020. 328 p.

и культуры. Но в реальности часто оказывается, что наборы данных перестают быть доступными, совместимыми или воспроизводимыми, как только конкретное исследование заканчивается, завершается проект или пропадает интерес у исследователя.

Для целей сохранения добытых данных создаются репозитории — специальные хранилища. В каждой предметной области создаются такие специализированные информационные системы. А нужно не забывать, что информационные системы — это не только серверы и программы, но и люди, их навыки, время и отзывы. В каких-то областях репозитории успешнее –например, корпусы в лингвистике или геоинформационные системы в истории, так как накопление данных общепризнано в сообществе. В других случаях сбор данных происходит медленно. Но суть требований к современным данным это не меняет. Если вы рассчитываете не только получить результат, но и сохранить данные для будущих поколений исследователей, не забудьте подробно задокументировать свои находки и подыскать для них репозиторий с хорошей репутацией.

В 1945 году Ванневар Буш написал статью «Как мы можем думать», которую сегодня многие считают пророчеством о будущей Всемирной паутине. Буш справедливо полагал, что обычной библиотеке не хватает синхронности с записями и сообщениями ученого, ради чего и нужно «проиндексировать» все необходимые для исследовательской работы тексты. Решать такую задачу должна машина, которую он назвал *Memex* (название объединяет два слова *memory* (память) и *index* (указатель)[1]. Очевидно, что с развитием современных интернет-технологий многообразие информации во Всемирной паутине создает подобный указатель зарегистрированных свидетельств памяти, которые обеспечены алгоритмами информационно-поисковых систем. Вопрос в том, чем электронные ресурсы отличаются от привычной библиотеки и есть ли у электронной библиотеки какие-то преимущества, помимо доступности и объема.

«На методологическом уровне я с огорчением наблюдаю, — пишет один из пионеров цифровых гуманитарных исследований Манфред Таллер, — что современное представление о цифровых инфраструктурах для гуманитарных наук, кажется, переоценивает идею публикации информации, поэтому инфраструктура для digital

[1] Bush V. As We May Think. 1945 July. URL: http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

humanities в ряде последних дискуссий может быть настолько лишена аналитических соображений, что становится практически неотличимой от цифровой библиотеки (и к тому же не очень сложной)»[1]. Это очень глубокое наблюдение.

По сути, речь идет о будущем научного поиска гуманитариев, способах построения новых исследований на основе данных (в международной литературе это называется data-driven research), а значит, данные — основа для новых форм научного творчества и интересных открытий.

[1] Таллер М. Дискуссии вокруг Digital Humanities // Историческая информатика. 2012. № 1. С. 11.

Глава 3

Культурное наследие и цифровые коллекции данных

(*И. А. Кижнер, М. В. Румянцев*)

За последние двадцать лет мы наблюдаем значительный рост количества оцифрованного контента в области культурного наследия и интереса к использованию цифровых материалов. Причины создания цифровых копий многообразны: необходимость управления коллекциями, демонстрация культурного наследия широкой публике, использование коллекций в научных, образовательных и творческих целях[1]. Часть цифровых коллекций культурного наследия составляют изображения (в отличие от оцифрованных текстов или аудиоданных)[2]. Рост количества опубликованных изображений начался в XVIII веке после изобретения литографии. Этот рост продолжился в геометрической прогрессии в XIX и XX веках, когда репродукция стала одним из базовых средств коммуникации. В XXI веке изменилось количество и качество цифровых изображений, увеличились объемы места хранения и появились сетевые технологии, позволяющие распространение изображений высокого разрешения[3], панорамных изображений, трехмерных моделей и иммерсивных технологий.

[1] Hughes L. M. (2004). Digitizing collections: strategic issues for the information manager (Vol. 2). Facet Publishing; Parry R. (2007). Recoding the museum: Digital heritage and the technologies of change. Routledge; Melissa Terras (2011). The Rise of Digitization. In Digitisation Perspectives, Ruth Rikowski. SensePublishers, Rotterdam, 3–20.

[2] Например, цифровая библиотека «Европеана» включает на момент написания этой главы около 28 миллионов изображений, или около половины из 58 миллионов учетных записей в коллекциях «Европеаны».

[3] Lynch C. A. (2002). Digital collections, digital libraries and the digitization of cultural heritage information. First Monday, 7(5).

Цифровые копии текстов и изображений составляют коллекции опубликованных документов. Такие коллекции можно найти на сайтах музеев, архивов и библиотек, но иногда они представлены как цифровые академические издания, опубликованные академическими сообществами или отдельными издателями. Цифровые копии в таких коллекциях сопровождаются метаданными (как описательными, которые включают данные о месте и времени создания объекта, так и административными, к ним относятся номера в реестрах и информация о месте хранения объекта в музее). Коллекции оцифрованных документов (текстов и изображений), опубликованных в цифровой среде, становятся частью инфраструктуры передачи знаний[1]. В рамках этого подхода цифровые коллекции воспринимаются как первичные источники, которые документируют особенности производства и распространения знания на разных этапах создания физических и цифровых коллекций.

Оцифрованные коллекции являются одним из источников данных для цифровых гуманитарных исследований. Иногда исследователи становятся инициаторами перевода физических документов в цифровую форму. После этого тексты или изображения становятся предметом дальнейшей обработки и анализа (например, текстового анализа или сетевого анализа). Для того чтобы превратить документы в цифровой объект, который может быть обработан машиной, необходимо сначала получить цифровую копию. Этот этап станет самым первым в ряду действий, которые затем приведут к получению набора текстовых или визуальных данных, а потом — к обнаружению закономерностей в большом наборе текстов или изображений.

Чаще всего созданием цифровых копий исторических документов и предметов культурного наследия занимаются музеи, архивы и библиотеки, используя внутреннее финансирование. Выбор текстов, предметов или изображений для оцифровки определяется разными политическими, экономическими и социальными причинами. Еще один важный фактор — научная повестка тех учреждений, которые проводят оцифровку, или тех исследователей, которые обращаются с просьбой предоставить цифровые копии текстов или предметов. В этих случаях выбор текстов или изображений определяется важностью или редкостью документов, а также востребованностью текстов или изображений в научном сообществе. Это может привести к тому,

[1] Mak B. (2014). Archaeology of a Digitization. Journal of the Association for Information Science and Technology, 65 (8), 1515–1526.

что та часть данных или коллекций данных, которая по какой-то причине не попала в фокус исследователей или широкой публики, так и не будет обнаружена из-за отсутствия цифровых копий[1]. Хотя цифровые копии существуют для объектов хранения многих коллекций музеев, архивов и библиотек, далеко не все коллекции оцифрованы, опубликованы и могут быть доступны исследователям для дальнейшего распространения[2].

Цифровые копии создают для предметов естественнонаучных коллекций и при оцифровке предметов культурного наследия[3]. Существуют правила и руководства по созданию цифровых копий для материалов культурного наследия, аудиозаписей и видеодокументов[4]. Однако более сложные культурные объекты, такие как клинописные таблички[5], наскальное искусство[6] или археологические объекты[7], могут подвергаться оцифровке с использованием техник трехмерного моделирования, таких как лазерное сканирование, фотограмметрия и другие сложные техники[8]. Предполагается, что

---

[1] Так, например, до сих пор не оцифрована и не опубликована как коллекция библиотека В. А. Жуковского, которая хранится в Томском государственном университете. Несмотря на то что каждая книга из такой коллекции в отдельности не является редкостью, это собрание признается важным датасетом для изучения источников распространения идей и интеллектуальных традиций в русском обществе первой половины XIX века.

[2] Terras M. (2022). Digital Humanities and Digitized Cultural Heritage. The Bloomsbury Handbook to the Digital Humanities / edited by James O'Sullivan, Bloomsbury Publishing. Pp. 255–266.

[3] Подробное описание рутинной оцифровки приведено в работе Thiers B. M., Tulig M. C., Watson K. A. (2016). Digitization of the new york botanical garden herbarium. Brittonia. Vol. 68(3). Pp. 324–33.

[4] https://www.digitizationguidelines.gov/

[5] Willems G., Verbiest F., Moreau W., Hameeuw H., Van Lerberghe K., Van Gool L. (2005). Easy and cost-effective cuneiform digitizing. In Short and Project Papers Proceedings of 6th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST2005), Mudge M., Ryan N., Scopigno R. Eurographics Association. Pp. 73–80.

[6] Mudge M., Malzbender T., Schroer C., Lum M. (2006). New reflection transformation imaging methods for rock art and multiple-viewpoint display. In: Ioannides M., Arnold D., Niccolucci F. Proceedings of the7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST2006). Eurographics Association. Pp. 195–200.

[7] De Reu J., Plets G., Verhoeven G., De Smedt P., Bats M., Cherretté B., De Maeyer W., Deconynck J., Herremans D., Laloo P., Van Meirvenne M., De Clercq W. (2013). Towards a three-dimensional cost-effective registration of the archaeological heritage. J. Archaeol. Sci. 40 (2). Pp. 1108–1121.

[8] D. Koller, B. Frischer, and G. Humphreys (2009). "Research challenges for digital archives of 3D cultural heritage models", in ACM Jounal on Computing and Cultural Heritage. Vol. 2. Iss. 3.

в этом случае трехмерная модель объекта культурного наследия имеет такую же информационную ценность, как и реальный предмет. Однако, как и всякая модель, трехмерная копия создается для того, чтобы выделить те особенности предмета, которые требуют пристального изучения. При создании модели отбрасывают «случайные черты». Нужно понимать, что те особенности, которые не учитывают при создании модели, могут оказаться важными при решении иной задачи или поиске ответа на иной вопрос исследования. В этом случае необходимо учитывать характер создания цифрового изображения, технические ограничения и контекст использования модели, поэтому трехмерная модель чаще всего имеет ценность в рамках конкретной, хорошо определенной цели, проекта или исторического контекста[1].

Цифровая копия часто определяется как произведение, имеющее культурную ценность (в отличие от ценности, которой наделяют уникальные или редкие предметы)[2]. Аутентичность или аура уникального музейного предмета — качества, которые трудно оценить количественным или качественным образом. Действительно, аутентичность иногда определяют как сочетание уникальности и исторической/культурной ценности[3]. В этом контексте культурная ценность цифровой копии может быть сопоставима с ценностью музейного предмета, вырванного из культурного ландшафта и таким образом утратившего часть своей оригинальности / исторической ценности. В процессе ремедиации часть культурной ценности и уникальности предмета неизменно теряется. Одновременно с этим приобретаются новые смыслы и новые культурные/исторические/социальные ценности. Приобретение новых смыслов или утрата прежних связаны с аннотированием цифровых копий музейных предметов и с той тематической коллекцией или классификационным разделом, куда попадает предмет в результате аннотирования.

[1] Hindmarch J., Terras M., Robson S. (2020). "On virtual auras. The cultural heritage object in the age of 3D digital reproduction". In Hannah Lewi, Wally Smith, Dirk vom Lehn, Steven Cooke. The Routledge International Handbook of New Digital Practices in Galleries, Libraries, Archives, Museums and Heritage Sites, Routledge.

[2] Meehan N. (2022). Digital museum objects and memory: postdigital materiality, aura and value. Curator: The Museum Journal. Vol. 65 (2). Pp. 417–434.

[3] Eberbach C., Crowley K. (2005). From living to virtual: Learning from museum objects. Curator: The museum journal. Vol. 48(3). Pp. 317–338.

Создание метаданных, или Аннотирование предметов культурного наследия

Традиционные подходы к оцифровке включают не только создание цифровых копий, но и аннотирование изображений, или работу по созданию метаданных[1]. Метаданные представляют собой структурированное описание предмета культурного наследия. Например, поля описания музейного предмета в базе данных могут включать такие разделы, как название, автор, дата, материал и техника, размеры, текстовое описание, названия выставок, в том числе в других музеях, где предмет появлялся в специальном контексте, подобранном куратором выставки, ссылки на научные публикации, в которых предмет упоминается. Структурированные данные дают возможность объединять часть данных в датасеты, организованные так, чтобы исследователь мог провести анализ и получить ответ на вопрос исследования. Такую работу можно автоматизировать, если музей, библиотека или архив предлагают воспользоваться интерфейсом прикладного программирования (API). В этом случае можно перенести большие объемы данных на стороннюю машину, с тем чтобы проводить исследования, используя одновременно метаданные и изображение предмета, добавляя дополнительные поля к описаниям, используя компьютерное зрение для анализа изображения или иные инструменты, необходимые для проведения исследования или работы над прикладным проектом. Возможность переноса данных, разумеется, возможна только, если музей, библиотека или архив сообщают правила, в рамках которых можно использовать данные. Иногда эти правила предполагают неограниченное использование части данных для произведений, созданных более ста лет назад (правила могут различаться в разных странах и для разных культурных учреждений). Однако в большинстве случаев использовать данные можно с большими ограничениями или нельзя использовать вовсе. Наличие больших объемов структурированных данных, для которых существуют явные правила, регулирующие использование в сторонних проектах, является важным условием работы в цифровых гуманитарных исследованиях. Еще одно важное

[1] Blagoderov V., Kitching I., Livermore L., Simonsen T., and Smith V. (2012). No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209: 133–146.

условие работы — совместимость метаданных, созданных в разных учреждениях и сообществах.

В литературе часто подчеркивается разнообразие стандартов метаданных в области описания предметов культурного наследия. Эта проблема приводит к тому, что данные сложно обнаружить, просмотреть, обеспечить их совместимость с другими данными и повторное использование в других проектах[1]. Последние два условия предполагают, что метаданные представлены в открытом некоммерческом формате, в машиночитаемой форме и их сопровождают лицензии, которые разрешают повторное использование метаданных. Можно выделить несколько подходов к аннотированию изображений, или работе с метаданными. Аннотирование изображений, или работа с метаданными, могут быть стандартизованы и сведены к минимальному количеству информации с обязательным включением идентификатора[2]. Альтернативный подход сводится к расширению кратких аннотаций, созданных на первом этапе оцифровки, с помощью автоматизированного аннотирования, основанного на машинном обучении, а также при участии специалистов и широкой публики в аннотировании коллекций[3]. Однако участие публики и привлечение волонтеров к аннотированию коллекций и переводу рукописного текста исторических документов в машиночитаемую форму оказывается трудозатратным процессом, который требует много времени, хотя и порождает высококачественные результаты, необходимые для анализа явлений культуры[4]. Поэтому наряду с оптическим распознаванием текста техникой, которая давно

[1] Poirier L., Fortun K., Costelloe-Kuehn B., Fortun M. (2020). Metadata, Digital Infrastructure, and the Data Ideologies of Cultural Anthropology. In: Crowder J., Fortun M., Besara R., Poirier L. Anthropological Data in the Digital Age. Palgrave Macmillan, Cham; Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N. et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data. Vol. 3(160018). P. 1–9.

[2] Blagoderov V., Kitching I., Livermore L., Simonsen T., and Smith V. (2012). No specimen left behind: industrial scale digitization of natural history collections. ZooKeys 209: 133–146.

[3] Hedrick B., Heberling M., Meinecke E., Turner K., Grassa C., Park D., Kennedy J., Clarke J., Cook J., Blackburn D., Edwards S., Davis C. (2019). Digitization and the future of natural history collections. BioScience. Vol. 70, Iss. 3. March 2020. P. 243–251.

[4] Weber A., Ameryan M., Wolstencorft K., Stork L., Heerlien M., Schomaker L.: Towards a digital infrastructure for illustrated handwritten archives. In: Loannides M. Digital Cultural Heritage. Information Systems and Applications, incl. Internet/Web, and HCI. Vol. 10605. Pp. 155–166. Springer International Publishing (April 2018).

и устойчиво применяется при работе с культурным наследием[1], но связана с ошибками и искажением смысла в процессе оцифровки и ремедиации[2], часто используется автоматизированное распознавание рукописного текста, основанное на техниках машинного обучения[3] и технике восстановления пропущенной информации на основе распознавания закономерностей в сходных объектах[4].

Пользователи цифровых коллекций культурного наследия

Кто же является пользователем цифровых коллекций культурного наследия? Кому нужны аннотированные цифровые копии текстов и визуальных материалов? В литературе показана значительная стратификация пользователей цифровых ресурсов в области культурного наследия и подчеркивается, что свыше 70% посетителей пользуются цифровыми коллекциями музеев не в учебных или образовательных целях, а в целях развлечения или в поисках вдохновения для создания собственных произведений искусства[5]. Рост количества посещений цифровых коллекций, до 28 миллионов

[1] Govindan V.K. and Shivaprasad A.P. (1990). "Character recognition — a review", Pattern Recognition. Vol. 23. No. 7. Pp. 671–683; Ul-Hasan A., Bukhari S.S. and Dengel A. (2016). "OCRoRACT: a sequence learning OCR system trained on isolated characters", 12th IAPR Workshop on Document Analysis Systems (DAS), IEEE. Pp. 174–179.

[2] Jarlbrink Johan and Pelle Snickars. (2017). Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive. Journal of Documentation 73 (6): 1228–1243.

[3] Terras M. (2005). Reading the Readers: Modelling Complex Humanities Processes to Build Cognitive Systems. Literary and Linguistic Computing, 20(1). Pp. 41–59; Muehlberger G. et al. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study, Journal of Documentation. Vol. 75. No. 5. Pp. 954–976.

[4] Lendemer J., Thiers B., Monfils A.K., Zaspel J., Ellwood E.R., Bentley A., LeVan K., Bates J., Jennings D., Contreras D., Lagomarsino L., Mabee P., Ford L.S., Guralnick R., Gropp R.E., Revelez M., Cobb N., Seltmann K. & Aime M.C. (2020). The extended specimen network: a strategy to enhance US biodiversity collections, promote research and education. BioScience. Vol. 70. Iss. 1. P. 23–30.

[5] Walsh D., Hall M.M., Clough P., Foster J. Characterising online museum users: a study of the national museums liverpool museum website. International Journal on Digital Libraries (Jul 2018); Elena Villaespesa (2019). Museum Collections and Online Users: Development of a Segmentation Model for the Metropolitan Museum of Art, Visitor Studies, 22:2, 233–252.

посещений в год для некоторых музейных ресурсов[1], может привести к дальнейшей стратификации пользователей, в том числе увеличению количества представителей общей публики. Такие пользователи предпочитают функциональность сайта, связанную с возможностью исследовать коллекции случайным образом, не прибегая к строке поиска и не используя специфические поисковые термины, особенно если это касается цифровых коллекций произведений искусства[2].

Интерес непрофессиональных пользователей, которым необходимы визуальные данные для собственной творческой деятельности, может быть обусловлен поиском культурной и этнической идентичности и онтологической защищенности[3]. Людям необходимо чувствовать принадлежность к разным формам культуры, аутентичность и авторитет которых поддерживают учреждения культуры. Им важно получить возможность творчества, соединенного с культурным наследием этническими, культурными и онтологическими связями. Более того, интерес к повторному использованию цифровых ресурсов связан с интересом к редкой в доцифровую эпоху способностью по-новому представить мир вокруг нас с помощью инфраструктуры, в которой можно легко распространить варианты репрезентаций[4]. Такая потребность в реконструировании и ремедиации цифровых ресурсов отвечает ожиданиям и нормам цифровой культуры[5].

Профессиональные исследования с использованием цифровых коллекций культурного наследия часто проводятся учеными, которые занимаются цифровыми гуманитарными исследованиями. В цифровых гуманитарных исследованиях опубликованные коллекции используются как источники данных. Такие исследования

---

[1] Ежегодные отчеты Музея Метрополитен в Нью-Йорке показывают увеличение количества посещений сайта от 700 000 в месяц в 2017 году до 2,5 миллиона посещений в месяц в 2019 году перед пандемией и резкий рост количества посещений в течение пандемии — до 28 миллионов в конце 2021 года (подробнее см. https://www.metmuseum.org/about-the-met/policies-and-documents/annual-reports).

[2] Lopatovska I., Bierlein I., Lember H. & Meyer E. "Exploring requirements for online art collections". Proceedings of the American Society for Information Science and Technology, 50(1), pp. 1–4.

[3] Turner B. (2001). 'The Erosion of Citizenship', British Journal of Sociology 52(2): 189–209.

[4] Couldry N. (2008). Mediatization or mediation: alternative understandings of the emergent space of digital storytelling. New Media & Society 10(3): 373–391.

[5] Deuze M. (2006). 'Participation, Remediation, Bricolage: Considering Principal Components of a Digital Culture', The Information Society 22(2): 63–75.

проводятся в области компьютерного анализа текста[1], исследования мультимодальных данных (сочетания текстов и изображений)[2], соотношения данных культурного наследия с культурными концептами пространства и географического местоположения с помощью географических информационных систем[3], анализа изображений[4], аудиоданных[5] и сетевого анализа[6]. Доступ к данным и возможности их анализа предоставляют разные учреждения культуры, исследовательские центры, библиотеки, музеи и архивы, такие как Исследовательский центр Хатхи консорциума учебных и научных библиотек университетов США[7], Центр предоставления данных

[1] Jockers M.L. Macroanalysis: Digital Methods & Literary History. University of Illinois Press, Champaign, Illinois, 2013; Underwood T. Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press. 2019.

[2] S. Janicke, G. Franzini, M. Cheema, and G. Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In Proc. of EuroVis — STARs, pages 83–103, 2015; Drucker J. (2014) Graphesis: Visual Forms of Knowledge Production. Cambridge, MA: Harvard University Press.

[3] Gregory I., Healey R. (2007). Historical GIS: structuring, mapping and analysing geographies of the past. Progress in Human Geography, 31, pp. 638–653; Paterson L.L., Gregory I.N. (2019). Geographical Information Systems and Textual Sources. In: Representations of Poverty and Place. Palgrave Macmillan, Cham.

[4] Besser H. (1990). Visual access to visual images: The UC Berkeley image database project. Library Trends 38(4): 787–798; Jörgensen C. (2003) Image Retrieval: Theory and Research.

Lanham, MD: Scarecrow Press; Melissa Terras. Digital Images for the Information Professional. Ashgate, Aldershot, England, 2008; Di Lenardo I., Seguin B., and Kaplan F. (2016) Visual patterns discovery in large databases of paintings. Proceedings of Digital Humanities 2016, Krakow. https://dh2016.adho.org/abstracts/348; Junginger P., Ostendorf D., Vissirini B.A., Voloshina A., Kreiseler S., Dörk M. Close-Up Cloud: Gaining A Sense Of Overview From Many Details. Presented at Digital Humanities 2019 Conference, 9–12 July, 2019.

[5] Müller M. Fundamentals of Music Processing; Springer International Publishing: Cham, Switzerland, 2015; Gref, Michael, Köhler, Joachim, Leh, Almut. 2018. Improved transcription and indexing of oral history interviews for digital humanities research. International Conference on Language Resources and Evaluation (LREC) 2018, http://publica.fraunhofer.de/dokumente/N-494202.html, 30 Jan 2019; Cord Pagenstecher 2019. Curating and Analyzing Oral History Collections. Selected papers from the CLARIN Annual Conference 2018. Linköping Electronic Conference Proceedings 159: 144–151.

[6] Knappett C. (2013). Network analysis in archaeology: New approaches to regional interaction. Oxford: Oxford University Press; Eder, Maciej. (2015). "Visualization in Stylometry: Cluster Analysis Using Networks." Digital Scholarship in the Humanities 30; Jackson C. (2017). Using social network analysis to reveal unseen relationship in medieval Scotland. Digital Scholarship in the Humanities, 32(2), pp. 336–343; Tamper M., Hyv¨onen E., Leskinen P. Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research. In: Proceeedings of CICLing 2019, Springer-Verlag (2019).

[7] https://www.hathitrust.org/htrc

Национальной библиотеки Нидерландов[1], Британская библиотека[2], Музей Метрополитен в Нью-Йорке[3], Отдел исторических газет Библиотеки Конгресса[4]. Несмотря на то что использование коллекций учреждений в области культурного наследия как источника данных осложнено вопросами авторского права людей и институций, которые участвуют в создании и ремедиации контента, опыт предоставления доступа к данным, разработки компьютерных средств обнаружения данных, их сортировки и анализа входит в практику библиотек, музеев и архивов[5].

Часто работа с метаданными приводит к тому, что исследователь замечает преобладание одного типа предметов и отсутствие других предметов или отсутствие нужных (ожидаемых) характеристик предметов. Например, документированная сеть людей, через которых Ганс Слоан приобретал предметы для своей коллекции[6], не включала представителей местного населения, тех, которые являлись источниками информации о предмете или доставляли сам предмет[7]. Часто это зависит от политики формирования музейных коллекций или традиций той эпохи, когда создается коллекция. В цифровых гуманитарных исследованиях отсутствие документации, которая объясняет причины включения предмета в коллекцию или особенности документирования предмета, становится важным ограничением использования коллекций. В этом случае классификация, анализ и сравнение осложняются метаданными, которые не объясняют забытые, скрытые, подразумеваемые мотивы включения или вовсе не объясняют место предмета в коллекции. Разные точки зрения на место предмета в классификационных схемах тоже осложняют анализ[8]. В последнее время в научной литературе часто поднимается вопрос экспликации (открытого, явного объяснения)

[1] https://www.kb.nl/en/resources-research-guides/data-services-apis

[2] https://www.bl.uk/catalogues-and-collections/digital-collections

[3] https://www.metmuseum.org/blogs/now-at-the-met/2018/met-collection-api

[4] https://chroniclingamerica.loc.gov/

[5] Allen L., Frost H., Padilla Th., Potvin S., Roke E.R., Varner S. (2019). The Collections as Data Framework: A Review from the Always Already Computational Project. https://tdl-ir.tdl.org/handle/2249.1/156364

[6] Коллекция Ганса Слоана стала основой коллекции Британского музея и Британской библиотеки.

[7] Ortolja-Baird A., Nyhan J. (2022). 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), 844–867.

[8] Kizhner I., Terras M., Afanasieva J., Pusenkova D., Sherer M., Skorinkin D. (2022). The culture of the very rich and very poor: Do museum digital collections tell us anything about Jewish culture? In Michelle Chesner, Amalia S. Levi, Daniel Stockl

причин преобладания одних данных и отсутствия других данных в больших коллекциях.

Открытый доступ к данным

Научные исследования в гуманитарных науках зависят от открытого доступа к данным в той же мере, в которой зависят от открытого доступа к данным и естественные науки. Это значит, что большие объемы открытых данных, которые хранятся в библиотеках, музеях и архивах, способны изменить то, как ведутся исследования, и способствовать развитию компьютерных инструментов для обнаружения закономерностей в данных. Концепция «открытого доступа», которой придерживается ряд учреждений культуры[1] для того, чтобы предоставить возможность повторного использования изображений из некоторых разделов своих коллекций, позволяет разным категориям пользователей копировать, изменять и распространять изображения, особенно те произведения, право копирования которых находится в общественном достоянии.

Концепция «открытого доступа» связана с возможностью получить бесплатный и беспрепятственный доступ к изображениям, а также с возможностью их использовать, изменять и распространять. Это определение следует формулировке, разработанной фондом «Открытое знание»[2] и применяется сообществом «Открытые библиотеки, музеи и архивы»[3]. Однако далеко не весь оцифрованный контент может быть доступен для повторного использования из-за ограничений, связанных с авторским правом и необходимостью лицензирования повторного использования. Право копирования творческого произведения переходит в общественное достояние через несколько десятков лет после смерти автора. Время, которое требуется для того, чтобы право копирования перешло в общественное достояние, регулируется национальными законодательствами.

Ben Ezra, Miriam Rurup, and Gerben Zaagsma. Jewish Studies in the Digital Age, Berlin, Boston: De Gruyter Oldenbourg, pp. 43–65.

[1] Kapsalis E. (2016). The Impact of Open Access on Galleries, Libraries, Museums, & Archives, Statewide Agricultural Land Use Baseline.

[2] http://opendefinition.org/

[3] https://medium.com/open-glam; McCarthy D. Uncovering the global picture of Open GLAM, Medium, 4 April, 2019. https://medium.com/open-glam/uncovering-the-global-picture-of-open-glam-af364aadeeee

Право копирования творческих произведений считается общественным достоянием и для работ, созданных в то время, когда авторское право еще не существовало. Однако часто учреждения культуры ограничивают право копирования и воспроизведения[1]. Это решение обосновывается тем, что созданная цифровая фотография тоже является предметом авторского права[2]. Учреждения культуры могут предоставить доступ к коллекциям через сайт музея, библиотеки или архива, ограничивая при этом использование изображений. В этих случаях использование данных в исследованиях оказывается затруднено. Это происходит из-за невозможности использовать инструменты нормализации данных, а также способы анализа или визуализации данных, которых нет на сайте, предоставляющем доступ к данным.

Особую проблему представляют произведения культуры и искусства, созданные в XX веке, которые не могут быть включены в программы массовой оцифровки. Этот пробел в создании и распространении знаний о культурном наследии часто называют «черной дырой двадцатого века». Публикацию и распространение знаний осложняет поиск правообладателя и бюрократические особенности лицензирования произведения для повторного использования, а в некоторых случаях полное отсутствие информации о правообладателе[3].

Большие коллекции данных

Копии оцифрованных предметов культурного наследия можно объединить в одну большую коллекцию, каталог или агрегатор. Для выполнения проектов такого типа необходимо объединить все источники данных, создать централизованную базу данных, обучить квалифицированных сотрудников и обеспечить долгосрочное

[1] Например, в соответствии с правилами, установленными Государственным Эрмитажем, необходимо получить письменное разрешение администрации музея для использования изображений в коммерческих и научных публикациях. См.: https://www.hermitagemuseum.org/wps/portal/hermitage/about/image_usage_policy/

[2] Pachali D. (2014). Open Access: How museums are opening their digital archives — Goethe-Institut. https://www.goethe.de/en/kul/bib/20365145.html

[3] Thylstrup N. B. The politics of mass digitization. The MIT Press, Cambridge, 2018.

финансирование[1]. Ряд хорошо известных агрегаторов метаданных в области культурного наследия, такие как Europeana Collections[2], DigitalNZ[3], Trove, агрегатор культурного наследия, связанного с Австралией[4] или Государственный каталог Музейного фонда Российской Федерации[5], были созданы по инициативе крупных государственных организаций, в то время как агрегатор Google Arts and Culture образовался по инициативе влиятельной коммерческой компании[6]. Особый акцент при создании таких коллекций делается на сопоставимости метаданных[7]. Однако цифровые ресурсы, представляющие культурное наследие, хранятся в учреждениях разного типа и в разных странах, что ведет к разнообразию стандартов описания и несовместимости моделей данных. Если библиотеки давно придерживаются общих стандартов метаданных, то архивы и музеи начали внедрять подобную практику совсем недавно. Проблема совместимости данных приводит к тому, что агрегаторы данных о культурном наследии, такие как, например, «Европеана», приводят исходные метаданные, полученные из учреждений культуры в разных странах, к единому формату и стандарту. Несмотря на то что этот процесс обеспечивает последовательное и совместимое представление данных, богатство и разнообразие исходных описаний и аннотаций пропадает в процессе конвертации[8]. Сочетание растущего объема цифрового контента, простых способов доступа с помощью метаданных, образовательных элементов и возможность использования данных для исследовательских и прикладных задач приводят к тому, что создание объединенных коллекций оказывается в фокусе представителей многих

[1] Lendemer J., Thiers B., Monfils A. K., Zaspel J., Ellwood E. R., Bentley A., LeVan K., Bates J., Jennings D., Contreras D., Lagomarsino L., Mabee P., Ford L. S., Guralnick R., Gropp R. E., Revelez M., Cobb N., Seltmann K. & Aime M. C. (2020). The extended specimen network: a strategy to enhance US biodiversity collections, promote research and education. BioScience. Vol. 70. Iss. 1. Pp. 23–30.

[2] https://www.europeana.eu/en

[3] https://digitalnz.org/

[4] https://trove.nla.gov.au/general/about

[5] https://goskatalog.ru/portal/#/

[6] https://artsandculture.google.com/

[7] Freire N., Voorburg R., Cornelissen R., de Valk S., Meijers E., Isaac A. Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network. Information 2019, 10, 252.

[8] de Boer V., Wielemaker J., van Gent J., Hildebrand M., Isaac A., van Ossenbruggen J., Schreiber G.: Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In: The Semantic Web: Researchand Applications, pp. 733–747. Springer (2012).

сообществ пользователей[1]. Еще одна функция агрегатора данных для предметов культурного наследия — обеспечить связь между метаданными и многочисленными справочными материалами, такими как данные о географических названиях, людях и событиях, с помощью уникальных идентификаторов, присвоенных предмету, и возможности объединять данные о явлении, событии или персоналии в пространстве связанных данных для того, чтобы установить связи между коллекциями[2].

Цифровые инфраструктуры, поддерживающие создание и распространение знаний[3] в области культурного наследия, помогают распространению и повторному использованию данных[4]. Примеры подобных коллекций включают цифровую библиотеку «Европеана»[5], Цифровую публичную библиотеку Америки[6], агрегатор цифровых ресурсов в области культурного наследия Новой Зеландии[7], платформу «Открытое наследие» Министерства культуры Франции[8], инфраструктуру связанных данных для изучения культурного наследия, объединяющую ряд университетов и организаций Канады[9], платформу Trove, которая предоставляет доступ к метаданным цифровых коллекций в области культурного наследия Австралии[10]. Некоторые из них дают возможность публикации произведений, право копирования которых находится

[1] Lynch C. A. (2002). Digital collections, digital libraries and the digitization of cultural heritage information. First Monday, 7(5).

[2] Julia Marden, Carolyn Li-Madeo, Noreen Whysel, and JeÂrey Edelstein. 2013. Linked Open Data for cultural heritage: Evolution of aninformation technology. In Proceedings of the 31stACM International Conference on Design of Communication. ACM, New York, NY, 107–112; Daquino M., Mambelli F., Peroni S., Tomasi F., Vitali F. Enhancing semantic expressivity in the cultural heritage domain: Exposing the Zeri Photo Archive as Linked Open Data. J. Comput. Cult. Herit. 2017, 10, 21–42; de Boer V., Wielemaker J., van Gent J., Hildebrand M., Isaac A., van Ossenbruggen J., Schreiber G.: Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In: The Semantic Web: Researchand Applications, pp. 733–747. Springer (2012).

[3] Mak B. (2014). "Archaeology of a Digitization." Journal of the Association for Information Science and Technology 65.8 (2014): 1515–26.

[4] Poirier L., Fortun K., Costelloe-Kuehn B., Fortun M. (2020). Metadata, Digital Infrastructure, and the Data Ideologies of Cultural Anthropology. In: Crowder J., Fortun M., Besara R., Poirier L. Anthropological Data in the Digital Age. Palgrave Macmillan, Cham.

[5] https://www.europeana.eu/en

[6] https://dp.la/

[7] https://digitalnz.org/

[8] https://www.pop.culture.gouv.fr

[9] https://lincsproject.ca/what-is-lincs/

[10] https://trove.nla.gov.au/

в общественном достоянии, в сторонних проектах. Действительно, цифровую библиотеку «Европеана» иногда определяют как портал, который в конечном итоге сможет дать возможность использовать большие объемы данных, предоставляя доступ с помощью прикладного интерфейса программирования (API)[1]. Открытое распространение данных способствует тому, что исследования приобретают дополнительную глубину и сложность, аудитория, которая пользуется данными, расширяется и стратифицируется[2]. Эпистемическая культура в цифровых гуманитарных науках часто требует моделирования данных и связей между данными, выделения тех черт в сложных объектах и произведениях культуры и искусства, которые могут быть исследованы цифровыми методами, и обнаружения связей между этими свойствами и внешним контекстом[3]. Цифровые коллекции дают возможность перехода от просмотровых и поисковых функций к аналитическим функциям и далее к функциям постановки задач и обнаружения закономерностей в данных.

Цифровые коллекции данных всегда конструированы из элементов доступных источников и отображают объективную картину только в некоторой степени. Оцифрованные документы, тексты и изображения не могут полностью отобразить масштаб, составные части и характеристики процессов, происходивших в культуре. Например, австралийская коллекция оцифрованных газет агрегатора Trove[4] включает только 30% газет, опубликованных в некоторые периоды XIX века. При работе с большими массивами данных в гуманитарных исследованиях лакуны в цифровых коллекциях и присущую им тенденциозность можно преодолеть с помощью

[1] Cesare Concordia, Stefan Gradmann, Sjoerd Siebinga (2010). Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. In: International Federation of Library Associations and Institutions 36(1), pp. 61–69 (http://dx.doi.org/10.1177/0340035209360764); Doerr M., Gradmann S., Hennicke S., Isaac A., Meghini C. & Van de Sompel H. (2012). The Europeana Data Model (EDM). Paper presented at the World Library and Information Congress: 76th IFLA General Conference and Assembly, Gothenburg, Sweden.

[2] Poirier L., Fortun K., Costelloe-Kuehn B., Fortun M. (2020). Metadata, Digital Infrastructure, and the Data Ideologies of Cultural Anthropology. In: Crowder J., Fortun M., Besara R., Poirier L. Anthropological Data in the Digital Age. Palgrave Macmillan, Cham.

[3] McCarty, Willard. Humanities Computing. London: Palgrave Macmillan, 2005; Flanders Julia and Fotis Jannidis. (2016). Data Modeling. In A New Companion to Digital Humanities, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 229–37. New York: Wiley & Sons.

[4] https://trove.nla.gov.au/newspaper/?q=

создания правильной выборки и сравнения данных[1]. Однако масштабирование исследований и применение техник статистического анализа не освобождают гуманитарную науку и исследования, связанные с изучением культуры, от ограничений, вызванных политическими, техническими и социальными причинами, которые приводят к недостаточной репрезентативности цифровых коллекций[2]. Необходимо, чтобы создатели и редакторы больших коллекций данных сообщали об ограничениях, лакунах и эпистемических пробелах, которые связаны с «историческими условиями, культурными и институциональными практиками, экономическими факторами и технологическими процессами»[3]. Деятельность по изучению репрезентативности и сбалансированности коллекции, способности прийти к объективным выводам на ее основе в результате анализа больших данных должна предшествовать количественным исследованиям и даже предоставлению данных широкой публике для распространения, изучения, создания новых продуктов и расширения цифрового канона. Призыв к явной и открытой демонстрации ограничений цифровых коллекций часто звучит в научном сообществе.

Заключение

Цифровые коллекции текстов, изображений, аудиоматериалов и видеофайлов предоставляют богатый материал для гуманитарных исследований. Метаданные и текстовые описания для таких коллекций публикуются квалифицированными исследователями с большим опытом изучения предметов культурного наследия в своей области. Авторитетность и качество данных/метаданных цифровых коллекций вызывают большой интерес научного сообщества. Несомненно, многое зависит от квалификации, образования и позиции людей, которые готовят метаданные и/или разметку текстовых данных. Важным обстоятельством является политический, социальный и экономический контекст создания физических и цифровых коллекций. Сочетание контекста, квалификации кураторов

[1] Underwood T. Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press. 2019.

[2] Bode K. (2020). Why you can't model away bias. Modern Language Quarterly. Vol. 81 (1). Pp. 95–124.

[3] Там же.

и обстоятельств оцифровки определяет те закономерности, которые увидят исследователи, готовые работать с большими наборами данных и использовать цифровые инструменты для извлечения и анализа изображений и текстов. Поэтому вопрос о том, как, где и почему проведена оцифровка коллекции, а также как, где и почему были документированы предметы культурного наследия, что было включено, а что опущено в метаданных, является важным вопросом при разработке проектов в области цифровых гуманитарных исследований[1]. Это обстоятельство становится еще более важным в эпоху, когда алгоритмы машинного обучения тиражируют самые видимые и частотные закономерности, обнаруженные в данных о культурном наследии.

Проблема воспроизводства только самых видимых и тиражируемых закономерностей приводит к необходимости понимания особенностей работы с общественно значимыми цифровыми данными. Образовательные программы обучения созданию цифровых данных в области культурного наследия становятся все более важными в контексте переноса навыков работы с цифровыми коллекциями в другие области (например, медицинские, образовательные или юридические наборы данных). В этом контексте важными являются не только практические технические навыки, но и знание причин, теорий и ограничений, сопутствующих созданию цифровых коллекций[2].

Развитие навыков использования и анализа цифровых данных происходит в среде, где пользователи оцифрованных данных получают необходимые инструкции и руководства по работе с культурно значимыми датасетами. Такие инструкции публикует ряд цифровых изданий, таких как Programming Historian[3] или Системный Блокъ[4]. Некоторые музеи, библиотеки и архивы начинают публиковать цифровые коллекции для использования без ограничений или с незначительными ограничениями. В таких коллекциях датасеты

[1] Zaagsma G. (2023). Digital History and the Politics of Digitization, Digital Scholarship in the Humanities. Vol. 38.2 P. 830-851; Terras M. (2022). Digital Humanities and Digitized Cultural Heritage. The Bloomsbury Handbook to the Digital Humanities, Bloomsbury Academic Publishing: London, New York, Dublin, pp. 255–267.

[2] Terras M. (2022). Digital Humanities and Digitized Cultural Heritage. The Bloomsbury Handbook to the Digital Humanities, Bloomsbury Academic Publishing: London, New York, Dublin, pp. 255–267.

[3] https://programminghistorian.org/

[4] https://sysblok.ru/category/courses/

сопровождаются обширной документацией, которая включает руководства по анализу данных[1].

Несомненно, существует необходимость взаимодействия между сотрудниками музеев, архивов и библиотек и исследователями, которые работают в области цифровых гуманитарных наук. Возможность взаимодействия, сотрудничества и передачи опыта будет работать в том и другом направлении. Этому могут способствовать совместные конференции, совместные проекты и публикации в изданиях, которые отличаются от тех, где обычно публикуются цифровые гуманитарные исследования[2].

Важно, что сотрудничество с музеями, архивами и библиотеками может проходить при создании и разработке проектов в области цифровых гуманитарных исследований. Не менее важно сотрудничать и для осознания ограничений, объяснения причин и создания теорий работы в цифровых гуманитарных науках. Только в этом случае распространение знаний о практических и технических сторонах формирования цифровых коллекций культурного наследия будет сопровождаться пониманием всех обстоятельств получения результатов исследований, обнаружения закономерностей в данных и развитии гуманитарного знания.

[1] См., например, соответствующий раздел цифровой коллекции Национальной библиотеки Шотландии: https://data.nls.uk/tools/

[2] Terras M. (2022). Digital Humanities and Digitized Cultural Heritage. The Bloomsbury Handbook to the Digital Humanities, Bloomsbury Academic Publishing: London, New York, Dublin, pp. 255–267.

Глава 4

Культуромика: исследование культуры и языка с помощью больших текстовых данных

(*А. А. Бонч-Осмоловская*)

В 2009 году Алон Хэлеви, Питер Норвиг и Фернандо Перейра, работавшие в это время в исследовательском отделе Google, публикуют статью-манифест The unreasonable effectiveness of data [Halevy et al., 2009]. Это статья была написана до «нейросетевой революции» конца 2010-х, в результате которой искусственный интеллект стал мейнстримом в науке и индустрии. Однако пафос и энергия статьи нисколько не противоречат, а напротив, предвосхищают будущее научно-технического развития. Статья начинается с противопоставления естественных наук, в которых для объяснения могут быть использованы математические формулы, и наук о человеке, которые, как пишут авторы, «резистентны» по отношению к стройной и прозрачной формализации. Альтернативой сложным теориям, которые никогда не будут обладать математической элегантностью, являются огромные корпусы текстовых данных,— «естественный союзник» наук, у которых в центре находится человек, а не элементарные частицы.

Культуромика является ярким и знаменательным примером такого союза, разворачивающим гуманитарные и социальные исследования в сторону больших данных. Ключевая идея культуромики состоит в том, что изменения частотностей слов рассматриваются как сигналы (или «следы) культурных и социальных изменений. Однако ценность культуромики состоит не просто в декларации квантитативного подхода, тем более что количественные исследования в гуманитарной сфере имеют длинную и продуктивную

историю с начала XX века. В основе культуромики лежит создание новой открытой исследовательской инфраструктуры беспрецедентного масштаба: датасетов корпусов текстов за несколько сотен лет на нескольких языках, подготовленных из текстов оцифрованных книг Google Books. Знаменательно, что Питер Норвиг, классик науки об искусственном интеллекте и один из авторов рассматриваемой выше статьи-манифеста, стал непосредственным участником этого проекта.

Эта глава будет организована следующим образом. Первые два параграфа будут посвящены описанию датасетов Google Books и особенностьям работы с ним с помощью инструмента Google Ngram Viewer, осуществляющему поиск и визуализацию результатов поиска. В первом параграфе рассмотрим, как был собран и создан датасет, определим его значимые характеристики и рассмотрим основные возможности на примере экспериментов, описанных его создателями в сопроводительной программной статье [Michel et al., 2011]. Во втором параграфе будут изложены основные принципы работы с инструментом Google Ngram Viewer. Далее, в третьем параграфе, рассмотрим основные аргументы критиков культуромики как метода и датасета Google n-грамм как исследовательского ресурса. В четвертом параграфе обсудим другие исследования, сделанные в научной парадигме культуромики, а также рассмотрим место культуромики как метода в общем контексте цифровых гуманитарных исследований.

1. Google Books как исследовательская база, культуромика как новый метод исследования культуры

В 2011 году в журнале *Science* выходит программная статья группы ученых под названием «Квантитативный анализ культуры с помощью миллионов оцифрованных книг» (Quantitative analysis of culture using millions of digitized books) [Michel et al., 2011]. В аннотации к статье авторы вводят новый термин — культуромика, который, по их мысли, должен дать название новому научному направлению — исследованию культуры с помощью текстовых больших данных.

Словообразовательная модель термина «культуромика» отсылает, конечно же, к «геномике», тем более что главные авторы статьи

много занимались биоинформатикой. Так же как геномика изучает совокупность генов в живых организмов, культуромика должна изучать некую совокупность уникальных единиц, из которых складывается культура. Такими единицами являются словоупотребления в гигантском массиве данных, состоящем из оцифрованных книг коллекции Google Books.

Ключевая идея статьи состоит в следующем: данные о том, как в течение времени меняется частотность определенных слов и словосочетаний, дают нам новые знания о социально-культурных трендах, общественных изменениях и процессах, открывает возможности сравнить и измерить такие, казалось бы, недоступные для квантитативных методов социальные концепты, как «популярность», «культурная память», «цензура и самоцензура», «внедрение технологий» и др. Эту идею авторы реализуют, во-первых, собрав огромный объем текстов разного времени и на разных языках, а во-вторых, разработав инструмент для поиска и анализа частотности слов в этих текстах, упорядоченных по году публикации, — Google Ngram Viewer.

История того, как сложился исследовательский коллектив, достаточно примечательна. Авторы изначальной идеи Эрез Эйден и Жан-Батист Мишель заканчивали аспирантуру в Гарварде и MIT в области биотехнологий: их диссертации не были связаны ни с лексикографией, ни с историей, ни с книговедением. Вот как они описывают в своей книге, почему они вообще начали думать о текстовых больших данных: «В течение длительного периода времени мы увлекались изучением истории. Особенно нас заинтересовал вопрос о том, как меняется со временем человеческая культура. Некоторые из этих изменений революционны, однако часто они оказываются совершенно незаметными для человеческого разума. Как было бы здорово, подумали мы, если бы в нашем распоряжении был какой-нибудь микроскоп для измерения человеческой культуры, выявления и отслеживания мельчайших изменений, совершенно незаметных обычному наблюдателю? Или же телескоп, позволяющий наблюдать с огромного расстояния — на других континентах или много столетий назад? Словом, возможно ли создать некий «скоп», помогающий наблюдать за историческими изменениями, а не физическими объектами?» [Эрец Эйден & Мишель, 2016].

Примерно в это же время Google начинает грандиозный проект по сканированию всех книг, которые когда-либо были изданы. Эйден

и Мишель понимают, что огромное количество текстов разного времени создания и поиск по ним и есть ключ к «скопу», который их интересует. Эрецу Эйдену удается встретиться с Питером Норвигом, директором по исследованиям Google, и презентовать ему идею создания «телескопа» по истории культуры. «Выслушав почти часовую презентацию Эреца, Норвиг наконец раскрыл свои карты. «Все это звучит прекрасно, но как мы сможем это реализовать, не нарушая закона об авторских правах?» [Эрец Эйден & Мишель, 2016: 82] Именно тогда возникает идея перевести связанные текстовые данные в датасеты последовательностей слов, связанных с годом публикации, n-граммы. Эйден и Мишель называют такое решение «тенью»: «Дело в том, что большие данные отбрасывают большие тени. Подобно тому, как тень представляет собой темную проекцию реального объекта — визуальную трансформацию, сохраняющую некоторые характеристики изначального объекта, при этом искажающую остальные, тень данных сохраняет часть изначальной информации» [Эрец Эйден & Мишель, 2016: 85].

Google Books удалось оцифровать порядка 15 миллионов книг, и это составляет примерно 12% от всех книг, которые когда-либо были опубликованы. Авторы проекта берут из этого массива треть — в датасет попадает около 5 миллионов книг, в которых меньше всего ошибок распознавания и нет проблем с метаданными. Метаданные очень важны, поскольку год издания книги является ключевой информацией, обеспечивающий возможность сравнения изменения частотности слов по годам. Полученный корпус, основанный на оцифрованных книгах, включал в себя тексты на семи языках — на английском (361 миллиард слов), французском (45 миллиардов слов), немецком (37 миллиардов слов), русском (35 миллиардов слов), испанском (45 миллиардов слов), китайском (13 миллиардов слов) и иврите (2 миллиарда слов).

В дальнейшем были увеличены и объем данных, и разнообразие корпусов: английский подкорпус был разделен на британский и американский английский, выделен отдельный подкорпус англоязычной художественной литературы (English fiction) и «миллионный корпус» английского[1]. К списку языков добавился корпус итальянского.

[1] Корпус Google one million является попыткой сделать сбалансированную по объему выборку на каждый год. Он устроен следующим образом: все книги, которые входят в корпус, написаны на английском языке и датируются от 1500 до 2008 года. Из одного года было выбрано не более 6000 книг, т.е. все отсканированные книги ранних лет присутствуют, а книги более поздних лет попали

Диапазон представленных данных огромен, самые ранние книги относятся к XVI веку, однако нужно понимать, что ранние данные не очень представительны, потому что в XVI веке издавалось очень мало книг, поэтому отдельные подкорпусы за год в XVI–XVIII веках очень небольшие. Только к XIX веку корпус Google Books вырастает до значительных величин — примерно до 100 миллионов слов в год. Именно поэтому авторы исследования советуют использовать данные начиная с 1800 года.

Корпус, созданный на основе книг, оцифрованных Google, нельзя читать подряд. Как отмечают авторы статьи, если попытаться прочесть только часть английского подкорпуса, начинающуюся с 2000 года, со средней скоростью 200 слов в минуту без перерывов на сон и еду, это займет 80 лет. Именно поэтому более точное название ресурса не корпус, а **датасет**, набор данных. В датасете хранятся не тексты и даже не отдельные предложения, а лишь последовательности слов длиной от 1 до 5. Такие последовательности называются **n-граммы**. Например, «Антарктида» — это униграмм, «Советский союз» — биграмм, «Китайская Народная Республика» — триграмм, а устойчивое словосочетание «The United States of America» — пентаграмм, сочетание из пяти слов подряд. Каждая книга была разбита на отдельные слова и словосочетания от 1 слова до 5 слов подряд, и для каждого такого слова или словосочетания было подсчитано количество употреблений: сколько раз это слово или словосочетание встретилось в тексте книги. Каждая книга имеет свой год выпуска. Таким образом, после обработки отдельной книги мы имеем список слов и словосочетаний, употребленных определенное количество раз в год, соответствующий году публикации книги. Далее эти данные обобщаются: для каждого года мы получаем список слов и словосочетаний с суммированной частотностью по всем книгам, изданным в этот год. В датасет попадают n-граммы, которые встретились во всем корпусе текстов не менее 40 раз. Формат представления книг с помощью частотностей n-граммов за каждый год позволяет выложить датасет в открытый доступ и дать возможность любому желающему использовать эти данные для собственных исследований. Таким образом, создатели датасета решают проблему авторских прав: сами тексты остаются недоступными для чтения и распространения, однако это и не нужно

в случайную выборку. Случайные выборки отражают тематические распределения по годам (так, в 2000 г. больше книг по компьютерам, чем в 1980 г.).

для метода культуромики, который оперирует изолированными частотностями слов и словосочетаний за каждый год.

Для того чтобы частотность слов в разные года можно было сравнивать между собой, используют механизм **нормализации**. Сравниваются не абсолютные значения частотностей, а относительные, скорректированные в зависимости от объема корпуса: частотность n-грамма делится на общее количество таких же n-граммов в подкорпусе за этот год. Одним из первых примеров, который приводится в статье, является динамика изменения частотности для слова slavery в американском корпусе английского (рис. 4.1). Для того чтобы сравнивать частотность этого слова, мы должны учитывать не только то, сколько раз мы его встретили в определенный год, но и как много вообще слов входит в датасет за этот год. Относительные частоты можно сравнивать и отображать в виде единого графика.

Необходимо сразу отметить, что «слово» в датасете Google n-грамм отличается от привычного лингвистического понимания и трактуется ближе к компьютерному термину «токен». Словом считается любая последовательность символов между пробелами. Соответственно, словами являются также сочетания букв и цифр, например «1-й», или просто цифры, например число π — «3.14159». Словами же будут признаваться и неправильно распознанные слова, слова, написанные с опечаткой, а также слова из другого языка. Эти данные не всегда являются шумом, которым можно пренебречь, часто они несут в себе важную информацию. Например,

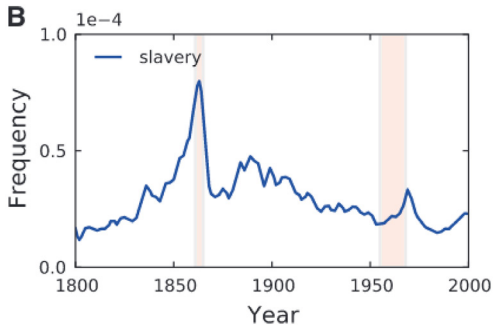


Рис. 4.1. Динамика частотности слова *slavery*. Мы видим два пика: один приходится на годы гражданской войны в США в 1861–1865 годах, другой — на годы борьбы против расового неравенства в 60-х годах XX века

один из экспериментов по исследованию динамики запоминания/забывания в культурной памяти, описанный в статье [Michel et al., 2011], связан с сравнением «жизненного цикла» чисел, соответствующих определенному году. Авторы сравнивают графики частотности чисел годов для каждого года от 1875 до 1975 (рис. 4.2) и обнаруживают, что, хотя все эти графики имеют одну характерную форму — резкий взлет в соответствующий год и потом медленное падение («забывание»), динамика изменения частотности (высота пика, скорость падения) меняется от XIX века к XX. «Мы забываем прошлое с каждым годом все быстрее» — к этому выводу приходят авторы статьи.

Таких «не-слов» достаточно много в датасете Google n-грамм. Авторы статьи делают примерную оценку доли английских слов для двух выборочных годов и получают результат в 51% от всех униграммов в 1900 году и 31% к 2000 году. Это изменение, по-видимому, говорит о том, что семиотическая система современного письменного языка оказывается шире собственно языковой и впитывает в себя множество иных неязыковых знаков, символов и паттернов. Однако и собственно «лексикон» датасета — то есть подмножество именно лексических единиц, очищенных от цифр и опечаток, — оказывается существенно больше объемов самых полных толковых словарей. На 2000 год количество лексических единиц в датасете

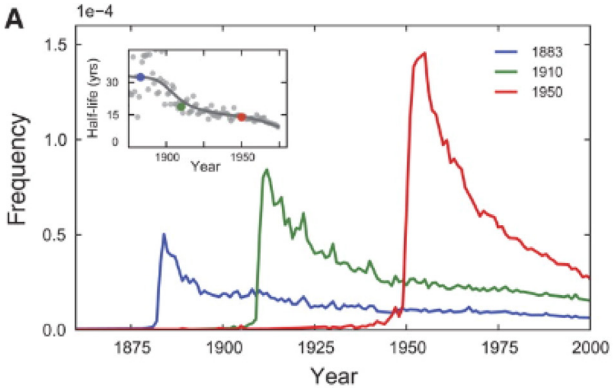


Рис. 4.2. Сравнение «угасания» памяти о годах. На маленьком графике показано сравнение длины периода, когда частотность упоминания года уменьшается вполовину после первичного «взлета». Можно видеть, что если в XIX веке такой период занимал 30 лет, то в середине XX века всего лишь 15

униграммов для английского языка составляет более 1 миллиона. Словник датасета, таким образом, оказывается существенно больше словников авторизованных лексикографических источников — словарей, которые должны документировать словарный запас языка. Авторы статьи предпринимают попытку исследовать то, что они называют «темной материей лексикона» (lexical “dark matter”). Они оценивают то, как увеличивается количество слов в английском датасете Google n-граммов с течением времени, и сравнивают эти объемы с количеством слов в наиболее авторитетных толковых словарях: с Оксфордским словарем, словарем Мериам-Вебстер и со словарем American Heritage Dictionary of the English language. Естественно, что для своего анализа авторы исследования исключают слова с опечатками, цифры или сочетания цифр и букв и имена собственные. Выводы из этого мини-исследования представляются очень интересными. Во-первых, как показывают авторы статьи, за последние 50 лет количество употребляемых слов увеличилось на 70% (рис. 4.3), и это очень резкий и показательный рост. Можно осторожно предположить, что он связан с тем, что изменилась индустрия публикации текстов в сторону существенно большей неформальности текстов и их тематического и, соответственно, лексического разнообразия. Кроме того, возможно, изменились и принципы отбора книг для библиотечного хранения, и в датасет стало попадать больше массовой или узкоспециальной научной

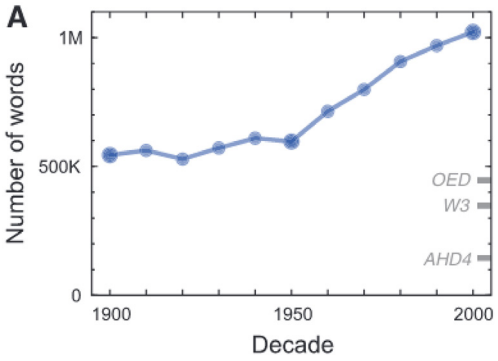


Рис. 4.3. Изменение количества «хороших» слов в датасетах Google n-граммов за XX век. Справа отмечено количество слов в авторитетных словарях: OED — Oxford English Dictionary, W3 — Miriam-Webster Dicitionary, AHD4 — American Heritage Dicitionary of the English Language

литературы. Об этом, кстати, говорится и в специальном исследовании датасета Google n-граммов [Pechenick et al., 2015]. Напомним, что книги были получены компанией Google для оцифровки из библиотек ведущих американских университетов, а это значит, что в датасете будет заведомо существенный перекос в сторону научной литературы по сравнению, например, с массовым популярным чтением. Из рис. 4.3 видно, что количество отдельных слов в английском датасете в 2000 году превышает более чем в два раза самый большой Оксфордский словарь английского языка. Авторы специально исследуют проблему, почему некоторые слова оказываются незаметными для традиционной лексикографии. Словари очень хорошо покрывают высокочастотные слова, но самые редкие слова туда не входят. Однако в соответствии с законом Ципфа[1] именно редкие слова составляют чуть более половины общего списка слов. Авторы оценивают «темную материю лексикона», которая не документирована ни в одном словаре, в 52% всего обобщенного словарного фонда английского языка.

Датасет, полученный из оцифровки книг Google, может быть использован для двух больших областей исследований. Во-первых, это лингвистические исследования изменения языка. Одно из них — оценка объема незафиксированного в словарях словарного фонда языка — было пересказано выше. В статье есть пример другого лингвистического диахронического исследования, посвященного оценке скорости нормализации неправильных форм прошедшего времени английских глаголов. Часть из форм прошедшего времени английских глаголов действительно постепенно нормализуется, иначе говоря, неправильные формы вытесняются продуктивной формой на *-(e)d*, другие глаголы, напротив, сохраняют нерегулярную форму прошедшего времени. Такого рода диахронические исследования достаточно широко распространены в корпусной лингвистике, и следует отметить, что, как правило, для них используются более тонкие статистические методы, чем просто сравнение

[1] Закон Ципфа описывает распределение частотности слов естественного языка: если слова корпуса текстов упорядочить по убыванию их частотности на шкале от самого частотного к наименее частотному, то частотность n-го слова в таком списке окажется приблизительно обратно пропорциональной его рангу в этой шкале. Например, второе по используемости слово встречается примерно в два раза реже, чем первое, третье – в три раза реже, чем первое, и так далее. Таким образом, распределение частотностей слов дает сначала очень быстро падающий вниз график, а затем длинный «хвост» из очень малочастотных слов.

частотностей[1]. Обычно исследуются «профили поведения» лексем и лексико-семантические конструкции, требующие более широкого контекста и более точных метаданных.

Более новаторским и в отношении методологии, и в отношении вводимых в научный оборот данных является исследование культурно-социальных трендов, разнообразные примеры которых представлены в рассматриваемой программной статье. Собственно, именно такого типа исследования и получают название «культуромики» — количественного измерения культурных трендов и их сравнительного анализа. Важно отметить, что, кроме собственно сравнения изменения частотности одного словосочетания или нескольких, в статье предлагается несколько более сложных подходов. Во-первых, используется метод, похожий на метод «семантических когорт» [Ryan Heuser & Long Le-Hac, 2012], когда сравниваются не отдельные слова, а группы слов, объединенные по некоторому смысловому признаку. В исследовании, представленном в статье, такими признаками стали профессии известных людей. Авторы исследуют феномен славы, ее пика и угасания, сравнивают тренды разного времени, а также то, насколько траектория известности (ее рост и падение) зависит от того, каким образом эта известность достигается, одинаков ли карьерный путь у известного математика и известного актера. Примечательно, что для такого исследования не хватает данных только n-граммов. Авторы берут из Википедии 740 тысяч имен, убирают оттуда людей с совпадающими именами и сортируют их по частоте упоминания и дате рождения. Для каждого года — с 1800-го по 1920-й — было выбрано 25 персон, родившихся в этом году, и ставших знаменитыми в одной из семи областей: актеры, художники, писатели, политики, биологи, физики и математики.

Далее было подсчитано, насколько часто представитель каждой профессии упоминался в зависимости от его возраста, а потом уже для профессиональных когорт в целом была вычислена медианная частотность упоминаний и ее изменения в зависимости от возраста обобщенного представителя когорты (рис. 4.4). Выяснилось, что быстрее всех — до 40 лет — известность приобретают актеры, 50 лет — критический возраст, когда политики становятся популярными, популярность писателей медленно растет в течение

[1] См. подробный обзор работ по диахронической корпусной лингвистике в [Hilpert & Gries, 2016].

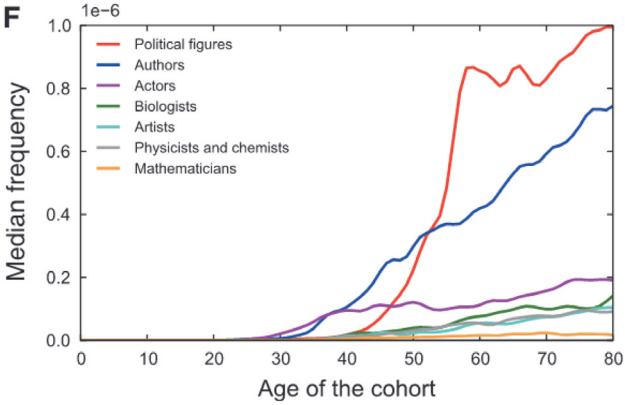


Рис. 4.4. Сравнение динамики роста известности для разных профессий

всей жизни, а известность математиков практически не связана с их возрастом.

Еще одним примером более сложного использования данных Google n-граммов является оценка так называемого индекса умолчания (suppression index). Авторы исследования пытаются найти подход, выявляющий то, как данные датасета отражают политико-социальные явления. Цензура — ключевой институт тоталитарных режимов, например, таких, как СССР и Германия в 1930–1940-х годах, — ограничивает свободную публикацию книг, а стало быть, датасет за эти годы должен быть искаженным. Можно ли с помощью метода культуромики обнаружить и показать подобное искажение? Самый простой способ — сравнение частотности упоминаний имен известных личностей, которые оказались под цензурным запретом в тоталитарной стране, но свободно упоминаются в публикациях стран без цензурных ограничений. Так, на наиболее, пожалуй, широко разошедшихся графиках, иллюстрирующих метод культуромики, демонстрируется резкий провал упоминаний Шагала в гитлеровской Германии (рис. 4.5) по сравнению с англоязычным датасетом, а также провал в частотности упоминаний Троцкого, Зиновьева и Каменева в русскоязычном датасете с 1936 года и вплоть до начала перестройки (рис. 4.6).

Несмотря на то что картина, показанная на рис. 4.6, представляется убедительной и соответствующей нашей интуиции, нельзя не отметить, что во-первых, в датасете невозможно отделить тексты,

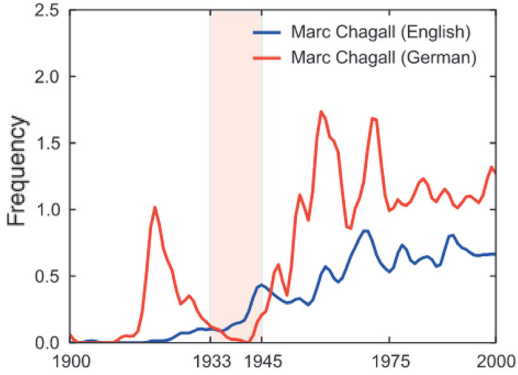


Рис. 4.5. Упоминание еврейского художника Марка Шагала практически сравнивается с нулем в конце 30-х годов в немецком датасете, несмотря на то что его популярность растет в англоязычном датасете

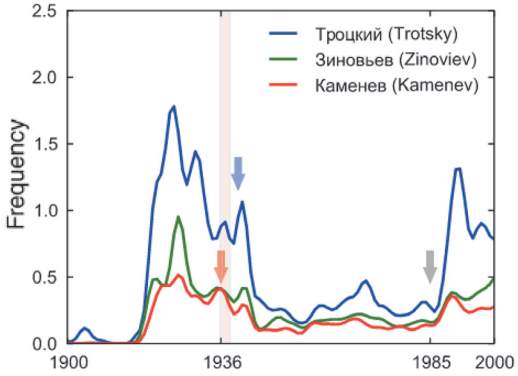


Рис. 4.6. Частотность упоминаний в текстах имен Зиновьева и Каменева падает вскоре после их расстрела в 1936 году (красная стрелка), частотность упоминания Троцкого резко снижается после его убийства в 1940 году (синяя стрелка). Действие цензуры, ограничивавшее упоминание этих политических деятелей, прослеживается вплоть до начала перестройки (серая стрелка), когда были сняты запреты и появились новые публикации, переосмысляющие прошлое

изданные в СССР, от текстов эмигрантских издательств, для которых не было цензурных запретов. Видимо, в частности поэтому упоминаемость Зиновьева и Каменева в советское время остается выше нуля, а для Троцкого мы даже видим некоторый подъем в 60-е

годы. Кроме того, отметим, что, в отличие от фамилии (псевдонима) Троцкий, фамилии Зиновьев и Каменев являются достаточно распространенными: только в Википедии можно найти не менее тридцати Зиновьевых и не менее десяти Каменевых, которые гипотетически могли быть упомянуты в рассматриваемый период времени. Это соображение имеет прямое отношение к критике культуромики как подхода, о котором подробнее поговорим в параграфе 3. Сейчас же заметим, что, возможно, более интересным и значимым, чем простое сравнение нормализованных частотностей по годам, являются обобщенные наблюдения над тем, какие искажения в распределении частотностей может давать цензура. Именно такой метод применяется при подсчете «индекса умолчания». «Умалчивание имен или идей оставляет свой след, который может быть посчитан» — пишут авторы статьи. Авторы берут списки деятелей культуры, политиков, философов нацистской Германии и дополняют этот список двумя контрольными списками — таким же списком, но для Великобритании, и списком деятелей национал-социалистической партии Германии этих лет. «Индекс умолчания» вычисляется как отношение средней частотности упоминаний имени из списка в период с 1935 по 1945 год к среднему от частотностей упоминаний этого же имени в 1922–1933 и 1955–1965 годах. Идея метода состоит в том, что если цензуры не было, то это отношение будет стремиться к единице: в течение рассматриваемых периодов известность человека скорее всего либо равномерно растет, либо равномерно падает, период в фокусе не выпадает из этого тренда, и поэтому средние значения будут примерно одинаковыми. Для англоязычного датасета мы получаем нормальное распределение, сконцентрированное вокруг единицы (синий график на рис. 4.7). Но для немецкого датасета картина другая: он очень сильно смещен влево, т.е. для большого количества имен из списка усредненная частота упоминаний в периоды с 1925 по 1933 год и с 1955 по 1965 год оказывается существенно выше частоты упоминаний с 1935 по 1945 год, а результат деления, таким образом, существенно меньше единицы. Но мы наблюдаем также и смещенные вправо результаты — это имена из списка деятелей нацизма, их упоминаемость, напротив, в периоды с 1925 по 1933 год и с 1955 по 1965 год оказывается существенно ниже, чем во времена Третьего рейха, а потому результат деления оказывается существенно больше единицы. Таким образом, кроме умолчания, мы можем наблюдать и другой (и, видимо, неотъемлемый от цензуры) тип искажения:

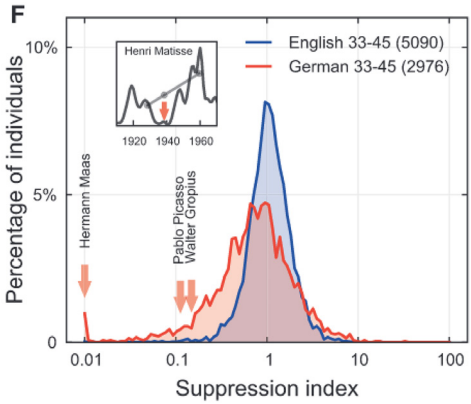


Рис. 4.7. Сравнение распределения индексов умолчания для выборки английского и немецкого датасета. Для английского датасета мы видим кривую нормального распределения Гаусса — «шляпу», которая концентрируется вокруг 1, это значит, что никаких существенных сдвигов в упоминаемости имен в период с 1935 по 1945 год не было и отношение упоминаемости имен в этот период примерно равно среднему их упоминаемости в предшествующий и последующий периоды. Для немецкого датасета мы видим смещенное распределение, на которое указывают красные стрелки. Красная область значений меньше единицы отражает существенно меньшее упоминание имен с 1935 по 1945 год, чем их среднее упоминание до и после этого периода. Самая крайняя точка, также отмеченная красной стрелкой, соответствует «индексу умолчания» Германа Мааса, протестантского священнослужителя, спасшего жизни многим евреям во время Холокоста. Во вставке также показан «индекс умолчания» для Анри Матисса, серая линия показывает линию роста его популярности с 1925 по 1965 год, а провал упоминаемости его имени с 1935 по 1945 год может быть объясним только внешним принуждением к его забвению, т.е. цензурой. Красная закрашенная область значений больше единицы соответствует именам деятелей нацистской партии, чей взлет упоминаний в период с 1935 по 1945 год является отзвуком тоталитарной нацистской пропаганды, частотность их упоминаний гораздо выше, чем ожидалось в целом, исходя из их известности в предшествующий и последующий периоды

«форсирование» или, как бы сейчас сказали, «накручивание» популярности деятелей правящего тоталитарного режима.

Итак, мы рассмотрели некоторые примеры культуромики как метода исследования, которые были представленные в статье [Michel et al., 2011]. С помощью этих примеров авторы демонстрируют диапазон исследовательских вопросов, которые могут быть поставлены исследователями датасета, а также некоторые методы работы

с данными датасета, в том числе более сложные, чем простое сопоставление частотностей слов и словосочетаний в разные годы. Датасет Google n-граммов может быть использован для лингвистических исследований, например, для оценки реального объема словаря и его изменений или для исследования диахронических процессов в грамматике и лексике на протяжении двух веков. Принципиально новое направление, для которого может быть использован датасет, — это измерение значимости изменений в частотности слов, отвечающих смене культурных трендов, отражающих влияние исторических событий и даже воздействие политических режимов. Датасет Google n-граммов дает возможность посчитать и сравнить самые абстрактные и, казалось бы, совершенно неквантифицируемые понятия, такие как «известность» или «коллективная память». Особая ценность работы, представленной в рассматриваемой статье, состоит в том, что ее результатом стали не только собственные исследования авторов, но и выложенные в открытый доступ уникальные датасеты, и самое главное — рабочий инструмент для дальнейших экспериментов по культуромике, доступный самой широкой аудитории, тем, кто интересуется культурой и хочет проверить гипотезы или наблюдения. Инструмент Google Ngram Viewer дает доступ к запросам по всем датасетам и не требует специальных навыков программирования для работы с ним. Ниже, в параграфе 2, мы рассмотрим в целом его основные функциональные возможности. Более подробно инструкцию по работе с Google Ngram Viewer можно изучить в специальном разделе на сайте https://books.google.com/ngrams/info.

2. Возможности поиска с помощью Google Ngram Viewer

Google Ngram Viewer — это инструмент для поиска по датасетам Google n-граммов. Он включает в себя строку для задания запроса, меню настроек запроса и экран визуализации результатов поиска в формате графика или нескольких графиков. В строке запроса можно задавать сразу несколько запросов через запятую, тогда результаты частотности каждого из запросов будут отображены на одном и том же экране визуализации с помощью графиков разного цвета. Такие графики очень удобно сравнивать. Меню настроек Google Ngram Viewer имеет следующие возможности:

– **Выбор временного периода с 1500 по 2019 год.** Как отмечают создатели инструмента, до начала XIX века, книг издается не очень много, поэтому данные этого периода не вполне представительны. Для русского датасета более или менее представительные данные начинаются с 1850 года.

– **Выбор датасета.** Выбор датасета включает в себя выбор языка и выбор версии языкового датасета.

– **Кнопка включения и выключения учета регистра**.

– **Кнопка выбора сглаживания**.

Иногда колебание результатов для разных годов мешает увидеть общую тенденцию. Для того чтобы увидеть результаты в обобщенном виде, используют механизм сглаживания. Уровень сглаживания задает диапазон результатов соседних лет с каждой стороны каждого года, по которым вычисляется среднее значение. Например, при уровне сглаживания 3 для 1980 года будет взято среднее для суммы значений 1977, 1978, 1979, 1980, 1981, 1982 и 1983 годов. Таким образом, вычисляется значение каждого года. Чем больше диапазон сглаживания, тем более обобщенной будет выглядеть линия тренда (рис. 4.8).

Кроме графиков, Google Ngram Viewer выдает ссылки на примеры в книгах, оцифрованных Google. Эти ссылки объединены во временные периоды, периоды поделены по количеству результатов на каждый год и поэтому могут быть совершенно неравномерными — если мы наблюдаем пиковые значения результатов, то временной период может включать в себя даже один год, а если результатов немного в течение большого временного отрезка, то тогда такой период объединяет десятилетия.

Несколько важных возможностей функционала Google Ngram Viewer связаны с языком запросов, который позволяет осуществлять поиск более гибко. Во-первых, в строку поискового запроса можно задавать сразу несколько запросов и сравнивать их результаты. Такие запросы пишутся через запятую (рис. 4.9).

Во-вторых, в запросах можно использовать **маску**. Знак * позволяет искать любое слово. Это может быть интересно для поиска словосочетаний, в которых одно слово заранее определено, а другое неизвестно. Результат поиска показывает 10 слов, которые чаще всего встречаются в заданной в запросе конструкции. Например, поисковый запрос с маской *советский и* * на диапазоне с 1918 по 2019 год (рис. 4.10) дает десять наиболее часто встречающихся слов, связанных со словом *советский* с помощью союза *и*: *постсоветский,*

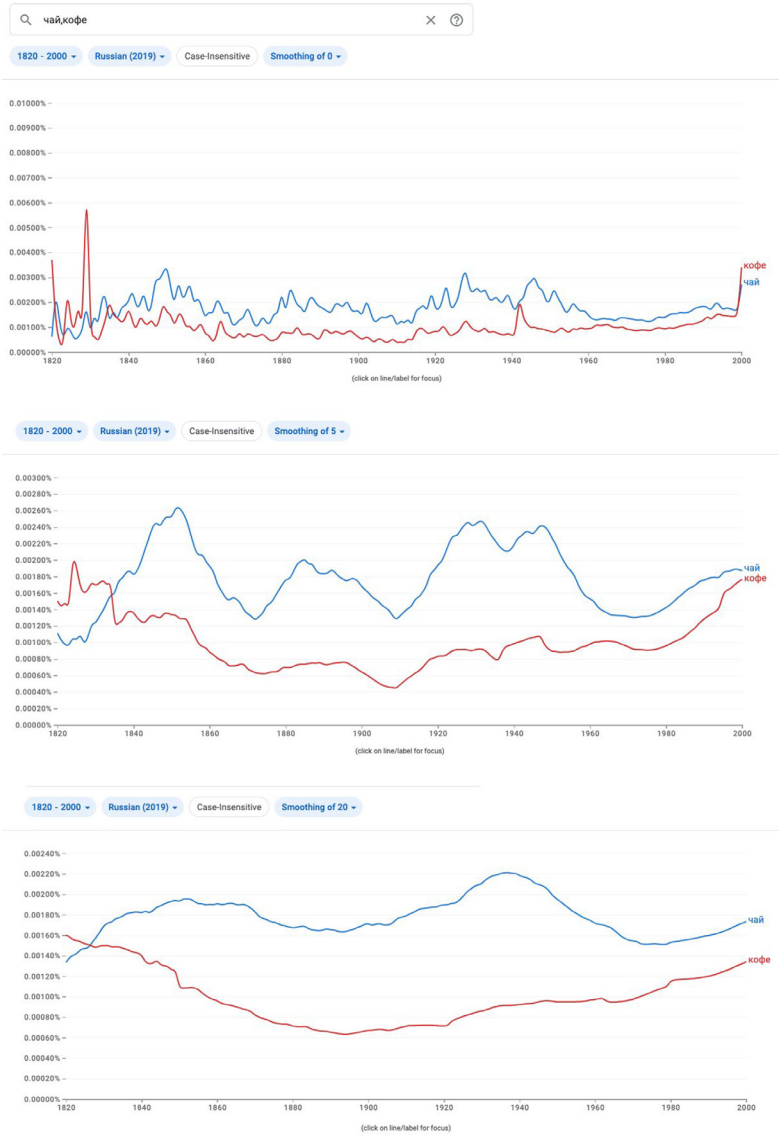


Рис. 4.8. Сравнение результатов запроса «чай, кофе» со сглаживанием 0,5 и 20. Отсутствие сглаживания показывает значения за каждый год. Сглаживание дает возможность увидеть самые значимые колебания и минимизирует случайные выбросы

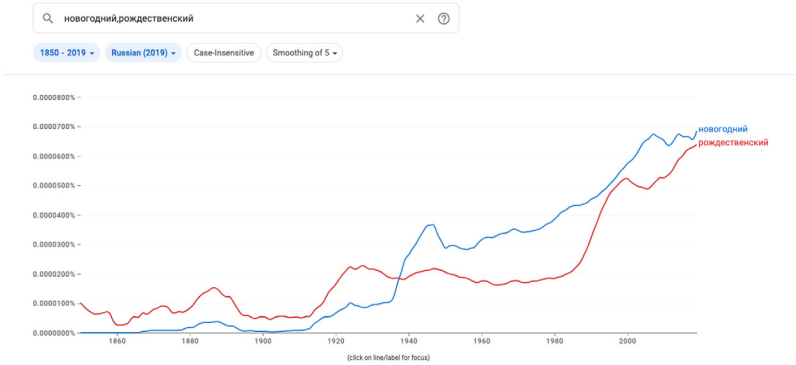


Рис. 4.9. Сравнение прилагательных *новогодний* и *рождественский*. Мы видим, что *рождественский* встречается чаще, чем *новогодний*, до середины 1930-х годов. В 1935 году празднование Нового года было официально разрешено, и прилагательное *новогодний* обгоняет и как бы вытесняет *рождественский* вплоть до середины 1990-х, когда празднование Рождества становится снова официально признанной практикой

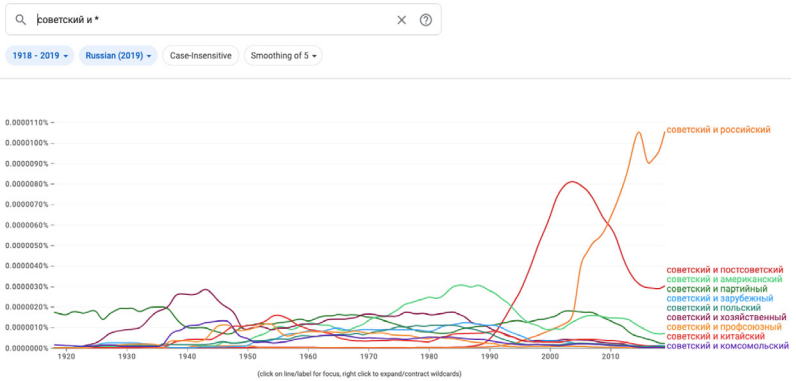


Рис. 4.10. Результаты поискового запроса с маской «советский и *». Наведение мышкой на конкретный график выделяет его и скрывает остальные, таким образом дает возможность лучше рассмотреть динамику изменений. На графике можно наблюдать «горбы централизации» в 1930–1940-х — *советский и хозяйственный, советский и партийный*, «дипломатические горбы» — *советский и китайский* в 1950-х, *советский и американский* в 1970–1990-х и «горбы исторической рефлексии» в 2000–2010-х — *советский и постсоветский, советский и российский*

российский, американский, партийный, зарубежный, китайский, польский, профсоюзный, хозяйственный, комсомольский. Маска дает нам возможность сравнивать не сами слова, но их контексты: изменение контекста употребления слова является очень хорошим сигналом культурных, социальных или языковых изменений. Так, например, на рис. 4.10 хорошо видны три типа контекста, взлет частотности которых отвечает трем историческим эпохам. Это, во-первых, сочетание слова *советский* с одним из прилагательных, обозначающих советские социальные или управленческие структуры, — *партийный, профсоюзный, хозяйственный, комсомольский*. Пики значений этих словосочетаний приходятся на 1930–1940-е годы, эти определения чаще всего сочетаются со словами *аппарат* или *актив*. Таким образом, мы видим «следы» централизации власти: формально разные компоненты социального устройства объединяются в общую структуру — советский и хозяйственный аппарат, советский и комсомольский актив. Еще один контекст слова *советский* — это сочетание с «географическим» прилагательным *американский, китайский, польский*, а также *зарубежный*. Такие контексты характерны для эпохи 1950–1980-х годов и отражают важные тренды во внешней политике СССР: дружбу с Китаем в начале 1950-х годов, разрядку в отношениях с США в 1970-х годах и в конце 1980-х, военное положение в Польше в начале 1980-х. Наконец, еще одна группа контекстов — *советский и российский*, *советский и постсоветский* — взлетает в частотности к началу 2000-х. Любопытно, что в начале 2000-х наиболее частотным является словосочетание советский и постсоветский, которое в преобладающем числе случаев сочетается с существительным *период*, однако примерно с 2010-х годов мы наблюдаем резкий взлет сочетания *советский и российский*. Можно осторожно предположить, что если сочетание *советский и постсоветский* несет в себе идею исторического перелома, то *советский и российский*, напротив, выражает идею исторической связанности СССР и Российской Федерации, которая стала чаще встречаться в текстах начиная с 2010-х годов.

Следует отметить, что поиск с маской имеет ограничение: маска может встречаться только один раз в словосочетании запроса; однако если через запятую записано несколько запросов, то у каждого из них может быть знак маски.

Еще одной возможностью языка запросов Google Ngram Viewer является возможность использовать в запросе теги частей речи, а также некоторые служебные теги. Это особенно важно для английских

датасетов, поскольку в английском языке слова часто не меняются формально при переходе из одной части речи в другую, и одно и то же внешне слово может быть и существительным, и прилагательным, и глаголом. Добавление к слову грамматического тега сужает запрос до определенной части речи (рис. 4.11).

Однако для русского языка грамматическая омонимия, подобная представленной на рис. 4.11, не характерна. Поэтому основной сферой использования грамматических тегов в русском датасете является их сочетание с маской: грамматический тег может накладывать ограничения на часть речи слова-маски и таким образом более жестко задавать структуру исследуемой конструкции. Так, на рис. 4.12 представлены результаты запроса, содержащего устойчивое словосочетание *Да здравствует* и заканчивающегося сочетанием маски и тега существительного *_NOUN. Такой запрос имеет своей целью выявить существительные, которые в разное время оказывались в контексте этого восклицания в советское время с 1918 по 1991 год. Полученные результаты достаточно интересны. Мы видим два «горба», в одном из них пиковые значения достигаются с середины 30-х годов до середины 40-х выражением *Да здравствует товарищ*, другой «горб» датируется серединой

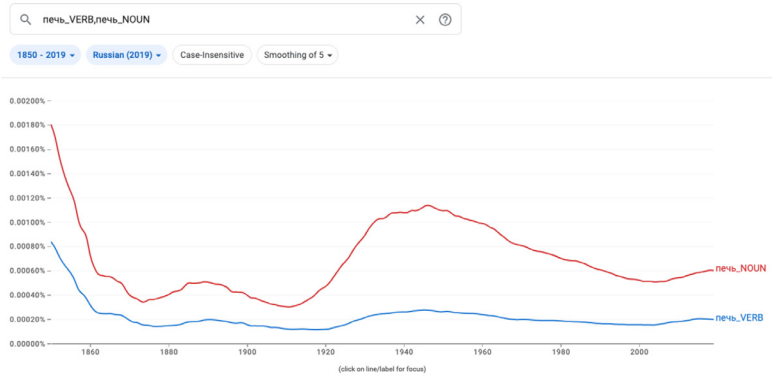


Рис. 4.11. С помощью грамматических тегов мы можем различить *печь*-существительное и *печь*-глагол и даже построить для них сравнительные графики. На графике видим «горб» для *печи*-существительного в середине XX века с пиком в военные годы. В это время особенно частотным является использование слова *печь* для описания процесса производства тяжелой промышленности — *плавильные печи, доменные печи, мартеновские печи*, тогда как глагол *печь* остается в зоне кулинарии

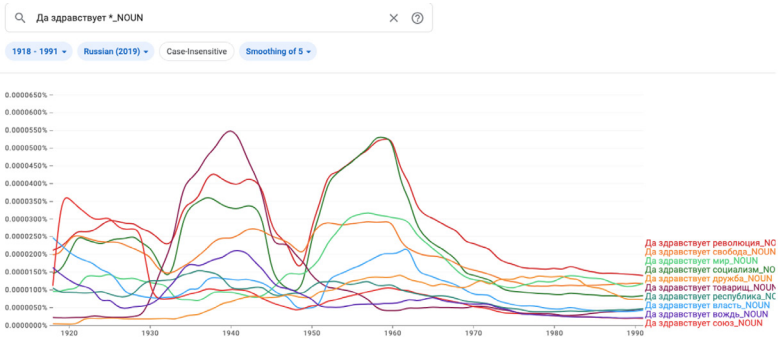


Рис. 4.12. Результаты запроса *Да здравствует *_NOUN* для периода с 1918 по 1991 год. Видны очень значительные колебания с 1930-х по 1970-е годы в частотностях разных выражений, два «горба» связаны с разными периодами жизни советского общества — эпохой культа личности 1930–1950-х годов и эпохой оттепели, выстраивания новой конфигурации международных отношений с середины 1950-х по начало 1960-х годов

1950-х — началом 1960-х годов, и тут максимальную частотность имеют выражения *Да здравствует революция*, *Да здравствует социализм, Да здравствует мир, Да здравствует дружба.* Следует отметить, что само выражение *Да здравствует X* — перформативное, т.е. говорящий, произнося его, одновременно осуществляет определенное речевое действие. Таким образом, это выражение должно относиться к чему-то, что происходит «здесь и сейчас» и что поддерживает и приветствует говорящий. Более пристальное изучение цитат, связанных с этими выражениями, показывает, что не все они являются истинно перформативными. Частотность выражения *Да здравствует революция* в середине 50-х годов связана с выходом большого количества исторических работ, посвященных революции 1905 года. И это выражение является прямой цитатой из листовок начала века, приводимых в этих работах. Другие частотные выражения этого же периода *Да здравствует мир, социализм, дружба* — перформативные и относятся к текущему моменту. Они связаны непосредственно со временем публикации, это в основном цитаты из официальных речей Хрущева и других партийных лидеров 60-х годов, времени оттепели, выстраивания послевоенного просоветского блока внешней политики, фестиваля молодежи и студентов в Москве. «Горб» 1930–1940-х годов можно считать «следом» эпохи культа личности. Хотя в выражении *Да здравствует товарищ* мы

не видим продолжения, сам период времени подсказывает, что полностью это выражение чаще всего звучит как *Да здравствует товарищ Сталин*. Действительно, если мы зададим уже более длинный запрос с маской на конце *Да здравствует товарищ **, результаты будут весьма выразительные (рис. 4.13). Для правильной интерпретации этого графика следует отметить, что слово или словосочетание попадает в датасет, если преодолевает порог в 40 употреблений [Michel et al., 2011: 176]. Таким образом, если запрос с маской выдает менее 10 результатов, значит, все остальные слова, употребляемые в этом контексте, не преодолели заданного порога минимальной частотности. Для выражения *Да здравствует товарищ ** мы получаем в ответ только четыре фамилии — *Сталин, Ленин, Молотов* и *Ворошилов*, это значит, что любые другие фамилии встречаются в этом контексте существенно реже и в датасет не попали. При этом трое из этого списка (Сталин, Молотов, Ворошилов) являются непосредственно деятелями эпохи сталинского культа личности, и лишь Ленин — «сакрализованная» фигура советского строя — не связан с конкретным временным периодом. Но обращает на себя внимание и фантастический разрыв между значениями частотности фамилии *Сталин* и остальными фамилиями. На рис. 4.13 приведены результаты запроса без сглаживания. Они показывают взлет частотности

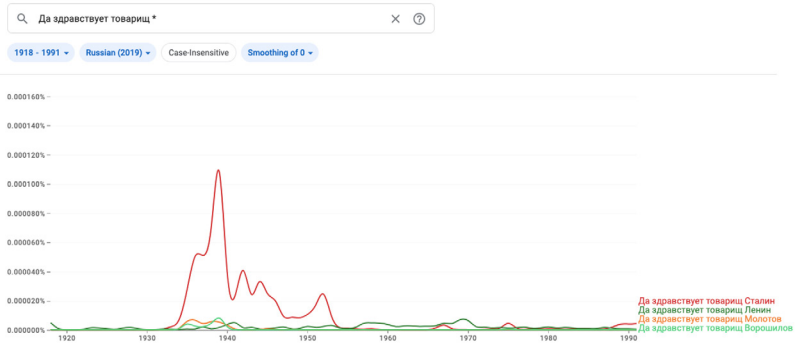


Рис. 4.13. Запрос с маской *Да здравствует товарищ ** выдает только четыре фамилии, все остальные имена встречаются в этом контексте слишком редко, чтобы попасть в датасет. График без сглаживания показывает более дробно, как через письменный язык выстраивался культ личности Сталина: практически вертикальный рост частотности в эпоху большого террора и «поддержание» культа в военные годы и в последние годы жизни Сталина во время послевоенного витка репрессий

выражения *Да здравствует товарищ Сталин* в 1930-е годы, в эпоху большого террора, и потом, после небольшого падения, характерную динамику роста в военные годы и в последние годы жизни Сталина. Иначе говоря, мы видим, как перформативное высказывание устной речи, относящееся не к абстрактному понятию, а к конкретному человеку, в определенный период начинает встречаться столь часто, что в разы превосходит по значению все остальные частотные слова в этом контексте. Именно метод культуромики и язык запросов Google Ngram Viewer дает возможность увидеть, как частотность формульных конструкций может быть связана со страшными периодами в истории нашего общества.

Язык запросов Google Ngram Viewer включает в себя еще целый ряд полезных функций, более подробно с ними можно ознакомиться на страницах соответствующего гайдлайна https://books.google.com/ngrams/info. Здесь упомянем некоторые из них:

- Возможность искать по всем формам слова по отдельности, а не только по обобщенной лемме. Для запроса используется тег INF, этот тег не сочетается в одном запросе с маской.
- Возможность искать не по последовательности слов, а по синтаксическим связям внутри словосочетаний. Для запроса используется знак стрелки => от главного слова к зависимому, сочетается также с маской и тегом части речи (рис. 4.14).
- Возможность объединять запросы в один с помощью скобок. Для русского датасета эта функция особенно важна, поскольку помогает учитывать написания в старой орфографии и объединять разные варианты написания в один запрос.
- Возможность «мультиплицировать» один из запросов. В том случае, если разница в частотности полученных результатов очень велика, график с низкой частотностью можно умножить на заданный коэффициент, таким образом, более видными и доступными для сравнения будут тренды изменений частотности двух графиков.

Итак, мы рассмотрели некоторые принципы работы с инструментом Google Ngram viewer, полное описание всего функционала представлено в разделе About Google Ngram Viewer на сайте https://books.google.com/ngrams/info, см. также подробное описание в [Захаров В. П. & Масевич А.Ц, 2014]). Однако использование Google Ngram Viewer — это не единственный способ работы с датасетами. Датасеты выложены в открытый доступ, их можно скачать и работать с ними напрямую.

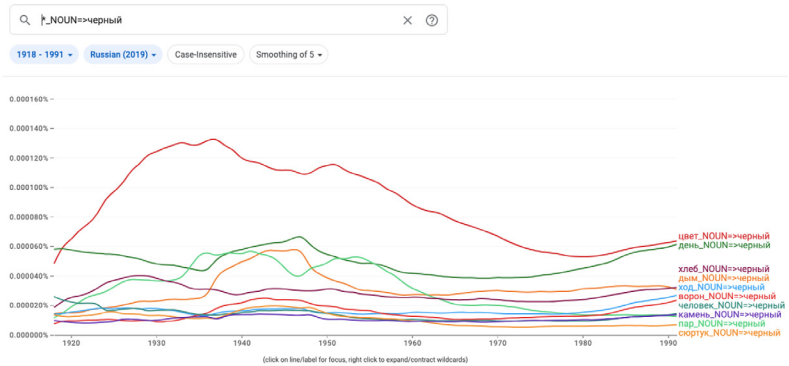


Рис. 4.14. Запрос, состоящий из маски, части речи, стрелки зависимости, выявляет наиболее частотные существительные, которые определяет слово *черный*. При этом *черный* может не стоять непосредственно рядом с определяемым словом. Запрос со стрелкой зависимости дает возможность уйти от требования задавать определенное количество слов в запросе и исследовать внутреннюю структуру лексико-синтаксических связей во всех n-граммах датасета

Публикация Google n-граммов — больших данных книг, оцифрованных Google, вместе со статьей в *Science*, демонстрирующей возможности работы с ними, имела очень широкий резонанс, как позитивный, так и негативный. Ниже мы рассмотрим основные аргументы критиков подхода, описанного в [Michel et al., 2011], и проблемы валидности датасетов Google Books.

3. Проблемные вопросы культуромики

Представляется очень важным отдельно остановиться на ограничениях, связанных с использованием датасета Google Books и культуромики как метода. С одной стороны, необходимо понимать технологические и концептуальные проблемы датасета, которые могут достаточно сильно исказить результаты и, соответственно, их интерпретацию. С другой стороны, дальнейшее развитие культуромики как метода во многом построено на преодолении этих проблем.

В целом, можно выделить три типа критических соображений, которые были высказаны в ряде публикаций в рамках дискуссии

о методе культуромики и связанных с ним ресурсах. Во-первых, это соображения технического характера, связанные с проблемами в подготовке данных датасета. Эти проблемы, в принципе, можно было бы учесть, если бы была поставлена задача заново собрать и обработать датасет. В целом, мы видим, что примерно раз в 5–7 лет датасет обновляется, туда добавляются новые данные, новая разметка и устраняются ошибки. Так что есть надежда, что в ближайшее время многие технические трудности будут преодолены.

Одна из ключевых проблем датасета — это недостаточное внимание к метаданным книг, из которых собираются n-граммы. В датасете никак не учитываются дубли и переиздания книг. А это значит, что иногда частотность какого-то слова или словосочетания может искусственно завышаться просто потому, что один и тот же текст присутствует в выборке несколько раз. Так, по всей видимости, произошло с выражением *Да здравствует революция* на рис. 4.12 — к 1955 году отнесены десятки книг практически идентичного содержания, посвященные празднованию революции 1905 года, и это обеспечивает пиковые значения этого выражения на этот год. Кроме того, при переиздании год издания книги считается годом употребления слова. Таким образом, например, юбилейное издание Пушкина или Толстого транслирует весь их вокабуляр в XX век. Решить эту проблему очень сложно, поскольку, как пишут авторы, проблемы с метаданными стали возникать уже на этапе сканирования. И собственно выбор 5 миллионов книг из 15 миллионов оцифрованных объяснялся именно тем, что только у этих книг более-менее в порядке были метаданные. Тем не менее представляется, что можно было бы автоматически сравнивать тексты книг и исключать повторы. Но проблема состоит не только в том, что повторяются тексты книг, но еще и в том, что сами метаданные — название книг, имена авторов и особенно названия издательств — попадают в датасет. Так, например, разница в результатах поиска словосочетаний *Московский рабочий*[1] и *Ленинградский рабочий* на рис. 4.15 объясняется тем, что *Московский рабочий* — это название издательства, основанного в 1922 году, и все книги, выпущенные в этом издательстве и попавшие в выборку Google Books, включают в качестве биграмма и его название. Разумеется, это является искажением, и эту проблему необходимо учитывать при поиске.

[1] Этот пример был предложен В.И. Беликовым для выявления проблемы с лишней индексацией метаданных в Google Ngram.

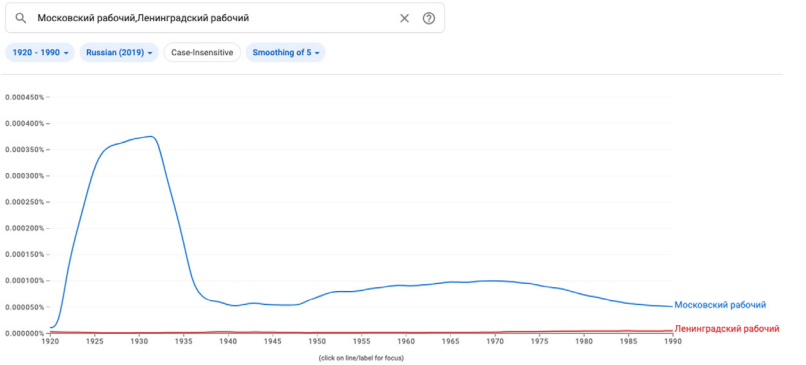


Рис. 4.15. Разница в частотности двух словосочетаний объясняется тем, что *Ленинградский рабочий* — это случайное словосочетание, а *Московский рабочий* — название издательства; таким образом, график отражает не частотность словосочетания как такового, а количество опубликованных книг в этом издательстве в каждый год

Еще одной часто встречающейся претензией к качеству датасета Google n-граммов является качество OCR — распознавание символов в тексте. Несмотря на то что эти проблемы последовательно исправляются, искажений все равно встречается достаточно. Например, существует проблема так называемого «длиного s» — архаичного написания латинского *s*, которое очень похоже на *f*. Достаточно задать пару английских слов, различающихся согласными *f* и *s*, чтобы увидеть, что в текстах XVIII века они оказываются практически неразличимыми (рис. 4.16 и 4.17).

Для русского датасета особую проблему представляет старая орфография, которая не совмещена с современной. И если *ъ* на конце или *i* еще можно учитывать в запросах (рис. 4.18), то слова с Θ (фита) или *ять* вообще не включены в датасет и не отображаются в запросе.

Как уже было сказано выше, проблемы технического характера могут быть решены постепенным улучшением датасета и корректировкой ошибок во входящих данных. Гораздо серьезнее проблемы концептуального характера, которые связаны с вопросами не к качеству подготовки датасета, а к методу и источнику данных. В общем, их можно сформулировать следующим образом: 1) насколько корректен метод сравнения частотности одного и того же слова или словосочетания в разные века и 2) насколько корректно

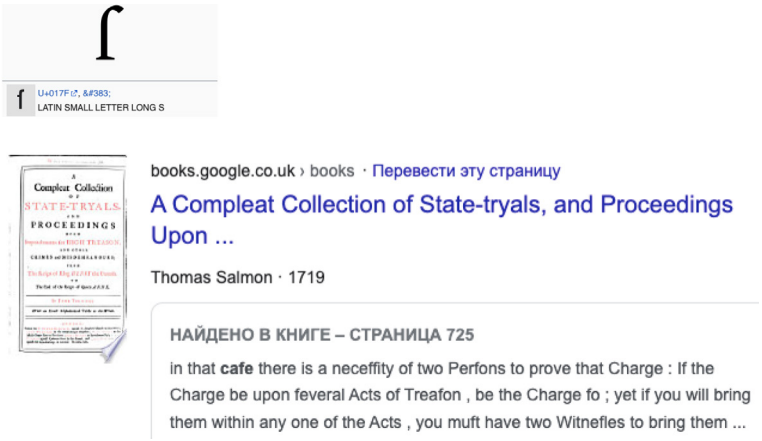


Рис. 4.16. Длинное *s* последовательно во всех словах распознается как *f*

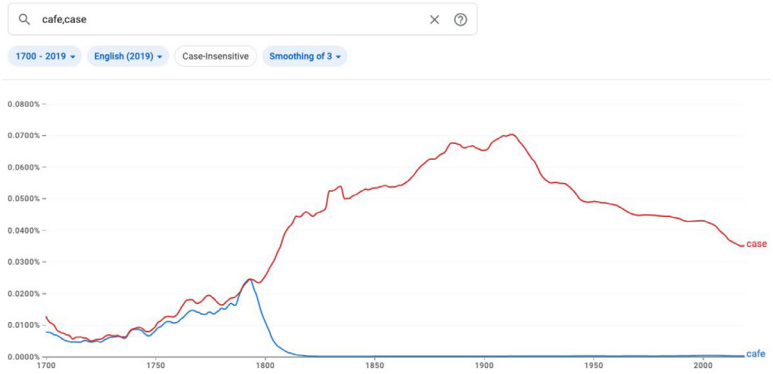


Рис. 4.17. Слово *cafe* гораздо менее частотно, чем слово *case*, однако в XVIII веке график динамики его частотности в точности повторяет график *case*, и это результат некорректного распознавания

делать выводы о культурных трендах, ничего не зная о тех текстах, из которых составлен датасет.

Проблема с уникальным словом как с «маркером» определенного культурного процесса состоит в том, что, в отличие от генома, язык избыточен и амбивалентен. А это значит, что, во-первых, слова могут получать новые и совершенно «посторонние» значения. Например, *мышка* в конце XX века начинает означать не только животное,

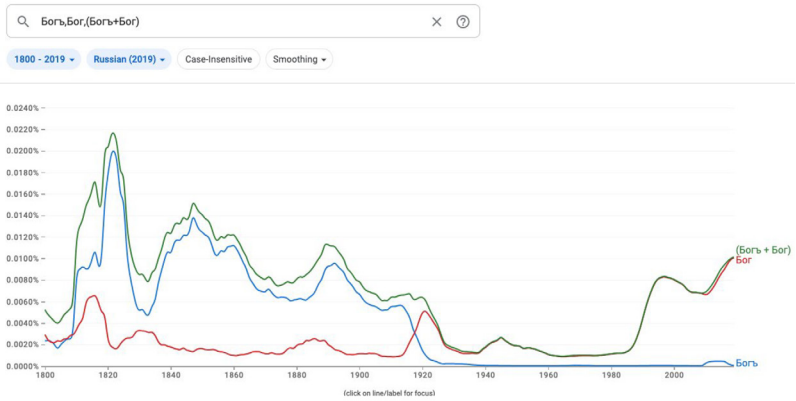


Рис. 4.18. С помощью объединения старого и нового способа написания слова (*Богъ+Бог*) можем получить график изменения частотности этого слова, отдельные запросы отражают именно частотность орфографического варианта, а не лексемы

но и приспособление для передвижения курсора на экране компьютера, а *ягуар* как марка машины в середине XX века практически полностью «забивает» употребление этого слова в его исходном значении названия животного. Во-вторых, значение слов в принципе может измениться. Так, например, слово *мотор* в начале XX века означало *автомобиль* [Даниэль & Добрушина, 2016], то есть, сравнивая частотность слова *мотор*, например в 1910-х годах и в 1990-х, мы сравниваем две разные сущности. Слово может полностью выйти из употребления и замениться другим — так, например, слово *рот* заместило слово *уста*. У слов, как правило, есть синонимы и сходные по смыслу выражения, и их все нельзя учитывать в одном запросе, тем более что они тоже меняются в течением времени. Все эти обстоятельства очень ограничивают поиск по Google n-граммам, и возможности культуромики как метода, заявленного в [Michel et al., 2011], в целом.

Но даже более существенной проблемой, чем лексические сдвиги, омонимия и полисемия, кажется проблема ограниченности и смещенности исходных данных датасета, а также недоступности источников текста для их фильтрации и более глубокого анализа наблюдаемых трендов. Как уже говорилось выше, Google Books изначально представляют собой собрание оцифрованных библиотечных книг. А это значит, что, проводя исследования на материале Google Books, мы

видим историю культуры и общества сквозь призму концепции наполнения университетской библиотеки. Таким образом, идея «популярности» преломляется достаточно непредсказуемым образом. Поскольку книги огромных тиражей и редкие издания представлены в библиотечных коллекциях одинаково небольшим количеством экземпляров, научный текст для специалистов и детективный роман массовой литературы будут рассматриваться как тексты сопоставимой доступности. Google Books никак не отражают известности текстов, мы не можем сказать, сколько человек потенциально могли прочесть тот или иной текст — десять или миллион, таких сведений нет и не может быть в цифровой библиотеке. Но именно поэтому выводы о росте или падении популярности того или иного слова или словосочетания могут быть весьма умозрительными: частотность упоминания свидетельствует лишь о том, что слово часто писали, но не о том, что его говорили или читали. В статье [Pechenick et al., 2015] показано, что в XX веке выборка Google Books оказывается перенасыщенной научной литературой, и это дает сильные сдвиги в частотных значениях относительно обычной лексики. Падение частотности обычных «бытовых» слов объясняется не выходом их из употребления, но их «растворением» в научной лексике. В статье [Pechenick et al., 2015], в частности, показано, как слово *Figure* с большой буквы, обозначающее график или рисунок в научной статье, с середины XX века оказывается гораздо более частотным по сравнению со словом *figure* с маленькой буквы, означающим «облик, изображение, фигуру…» (рис. 4.19); и это, как убедительно показывают авторы статьи, результат смещенной выборки, а не реальное положение вещей.

Американский ученый, специалист по digital humanities Тед Андервуд в своем блоге[1] замечает, что даже работая только с датасетом художественной литературы, мы все равно ничего не знаем о том, как распределяются жанры внутри датасета, и таким образом не можем отделить собственно культурные изменения от литературного процесса. Андервуд приводит пример со словом *fidelity* (верность), частотность которого неуклонно падает в датасете English fiction с начала XIX века. Свидетельствует ли это об изменившихся нравах или же просто о вытеснении жанра сентиментального романа, в котором это слово было сверхчастотным — спрашивает Андервуд.

[1] https://tedunderwood.com/category/ngrams/ (доступно 20.06.2023).

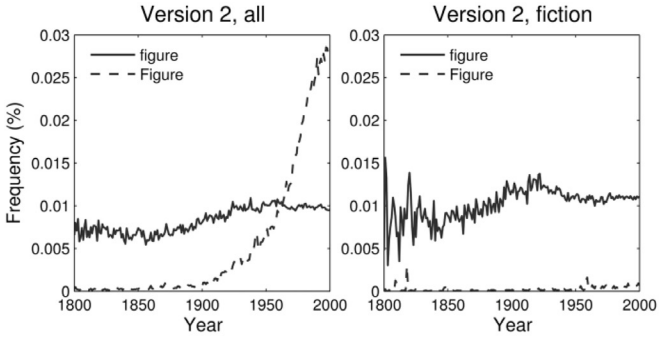


Рис. 4.19. В статье сравниваются два графика: первый график отражает вторую версию датасетов (2012) общего англоязычного датасета и отдельного датасета художественной литературы. По взлету частотности *Figure* с большой буквы в общем датасете мы можем судить о том, насколько сильно сдвинут в сторону научной литературы датасет в XX веке относительно XIX века [Pechenick et al., 2015]

Лингвист Марк Либерман рассуждает в своем блоге[1] о недостаточности данных Google n-граммов. Он отмечает, что, к сожалению, некоторые вопросы нельзя свести только к частности слов, для них нужны другие типы анализа, которые не обеспечиваются датасетом. Если сравнивать создание Google n-граммов с расшифровкой генома человека (а на такое сравнение намекает само название метода), то ключевая проблема, подчеркивает Либерман, состоит в том, что у нас нет доступа к источникам. Даже тексты, не связанные с авторским правом, — это собственность частной компании Google, вложившейся в их оцифровку и распознавание, и едва ли кто-то в ближайшее время сможет повторить эти усилия. Расшифровка генома человека была общественным, а не частным проектом. Информация, которую мы получаем сейчас из датасетов Google n-грамм, эквивалентна знанию об относительной распространенности отличий последовательностей ДНК среди индвидуумов в геноме, без доступа ко всему геному как таковому.

[1] https://languagelog.ldc.upenn.edu/nll/?p=2848

4. Культуромика как научное направление

Статья [Michel et al., 2011], а в особенности инструмент поиска n-граммов в книгах Google Ngram Viewer имели очень широкий резонанс, выходящий за пределы академических изданий. Заметки и колонки про исследования культуры сквозь призму больших данных книг вышли в таких общественных и научно-популярных изданиях, как Wall Street Journal, The Gaurdian, Wired, Scentific American и в многих других [Shea, 2012; Jha & Kingsland, 2010; Zhang, 2015; Harmon, 2010]. Таким образом, культуромика воспринималась и позиционировалась в большей степени как научно-развлекательное упражнение. На фоне критических откликов от академического сообщества было не совсем понятно, насколько поиск и интерпретация Google n-граммов могут войти в серьезный научный аппарат. Спустя более чем 10 лет после выхода статьи мы можем утверждать, что идея «дальнего чтения» культуры с помощью анализа данных гигантских корпусов текстов получила свое развитие в академических исследованиях digital humanities. Можно говорить о трех направлениях развития культуромики как научного направления. Это, во-первых, исследования, проведенные на данных датасета Google Books, однако с привлечением дополнительной специализированной информации, благодаря которой наблюдаемые графики получают более глубокую интерпретацию. Во-вторых, это исследования, которые используют идею культуромики на других датасетах, как правило, не содержащих проблем, о которых говорилось в параграфе 3. И наконец, это сформировавшаяся новая область знаний в области экологии и охраны природы (conservation culturomics).

Научные исследования с помощью Google n-граммов, как правило, строятся как серия исследовательских запросов к датасету, каждый из которых расширяет и усложняет понимание о культурных и языковых изменениях. Интерпретация, а также подбор ключевых слов запроса, как правило, опираются на дополнительные внешние источники. Таким образом, если в статье [Michel et al., 2011] мы видели множество отдельных интересных экспериментов, иллюстрирующих возможности метода и ресурсов, специализированные исследования по культуромике на датасете Google Books фокусируются на одной теме, получая в итоге более сложные и многогранные результаты. В статье [K. Willems, 2013] исследуется язык Третьего рейха с помощью датасета Google n-граммов. Автор

проверяет гипотезу о том, что лишь небольшое количество выражений и словосочетаний нацистского режима были сформированы в это время. Большинство из них существовали и ранее, но были переосмыслены в рамках нацистской идеологии и стали существенно частотнее. Входными данными для исследования являются 50 выражений из словаря языка Третьего рейха. Оказывается, что 40 из них использовались до 1918 года, часть из них сейчас практически не употребляется, поскольку коннотации с нацистским периодом оказались слишком сильны (рис. 4.20), другие, наоборот, потеряли

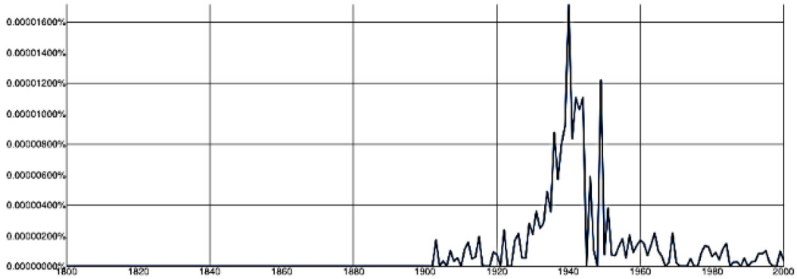


Рис. 4.20. На графике показано, как архаичное слово *Meintat* (преступление) становится частью идеологического языка Третьего рейха и резко растет в частотности, а потом так же резко падает после окончания Второй мировой войны, оставаясь окрашенным «нацистским» словом [K. Willems, 2013: 7]

Рис. 4.21. График иллюстрирует изменения в частотности немецкого слова *Großoffensive* (широкомасштабное наступление), которое, как утверждает Словарь языка Третьего рейха [Michael, Robert and Karin Doerr, 2002], использовалось Геббельсом, чтобы поднять боевой дух немецких солдат. Однако это слово остается частотным в немецком языке и после Второй мировой войны, теряя свои коннотации с языком нацистского режима [K. Willems, 2013: 10]

связь с языком Третьего рейха и продолжают использоваться в нейтральном значении до сих пор.

Еще одним интересным примером глубокой интерпретации Google n-граммов является работа [Ophir, 2016]. Автор берет понятие *truth* (правда, истина) как основное слово для исследования и выстраивает последовательную цепочку запросов. Сначала задается запрос с маской *truth**, наиболее частотным ответом на запрос является сочетание *truth and* (правда и). На следующем этапе этот наиболее частотный ответ становится запросом с маской *truth and* *. Далее строится сложный запрос, в котором через запятую указываются 10 наиболее частотных слов из ответа на предыдущий запрос, т.е. такие, которые чаще всего встречаются в сочинительной конструкции со словом *truth*, из них выбираются два, графики которых лучше всего коррелируют со словом *truth* — *justice* (справедливость) и *love* (любовь) (рис. 4.22).

Продолжая использовать метод последовательных уточняющих запросов, автор рассматривает корреляцию *Truth* и *Love* и показывает, что конструкции *Truth of* и *Love of* наиболее часто встречаются вместе с существительным *God* (Бог). Для теоретического обоснования наблюдаемых флуктуаций автор привлекает работу американского социолога русского происхождения Питирима Сорокина. Сорокин создал уникальную классификацию шести философских направлений и их понятийных паттернов и показал смену и взаимодействие этих направлений на протяжении 2500 лет. Исследования

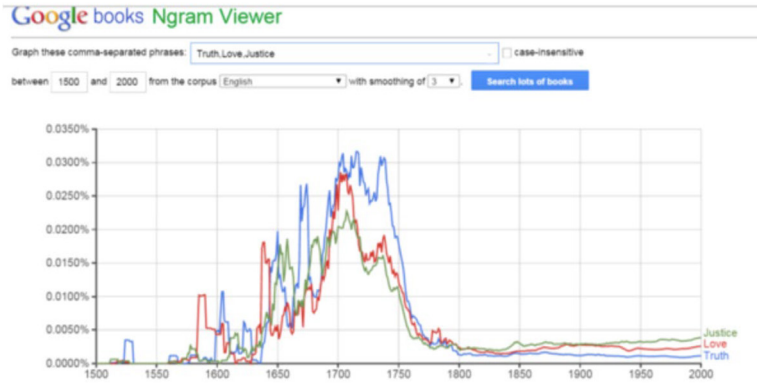


Рис. 4.22. График выявляет корреляцию между словами *Truth, Love, Juctice* с пиком значения частотности употребления в начале XVIII века [Ophir, 2016]

Сорокина были проделаны вручную и опирались на анализ текстов и на внешние факторы. Два паттерна — «этика любви» и «системы истин» — связываются Сорокиным как элементы *Рационализма*, философии идей, достигающего своего пика в начале XVIII века. Это фундаментальное теоретическое исследование очень хорошо соотносится с результатами исследования, проведенного на материале датасета Google Books.

Доминирующее большинство современных исследований культуры с помощью анализа диахронических текстовых данных не ограничиваются только Google n-граммами, а иногда и вообще используют другие корпусы текстов. Безусловно продуктивной оказалась сама идея совмещения числовых метрик частотности слов и словосочетаний и временной шкалы как способ получения объективных и квантифицируемых показателей изменений значимости культурных феноменов. Можно утверждать, что дальнейшая эволюция культуромики как метода состоит в поиске способов «объективизации» цифровых данных. Большинство решений, которые предлагаются, направлены на преодоление «родовых болезней» датасета Google Books, о которых говорилось выше (в параграфе 3), например, используются новые подходы к нормализации данных, чтобы минимизировать неравномерную представленность текстов в разное время, расширяются лексические паттерны для поиска, чтобы включить синонимию, используются корпусы с более диверсифицированными метаданными и жанровыми системами.

Статья [Morin & Acerbi, 2017] посвящена исследованию эволюции эмоционально нагруженных выражений в английской художественной литературе. Они показывают, что в течение XIX века снижается эмоциональная лексика с позитивным значением, но растет негативно окрашенная лексика. Датасет Google n-грамм корпуса английской художественной литературы (English fiction) используется в качестве базового ресурса, наблюдения, полученные в результате анализа эмоционально нагруженных слов на этих данных, в дальнейшем проверяются на двух небольших корпусах, которые имеют хорошую разметку метаданных. Один из этих корпусов — это корпус классических произведений британской литературы, другой корпус составлен из достаточно случайного набора книг, к которым имелся открытый доступ. Корпусы не пересекаются по содержанию, сбалансированы по объему, количеству представленных авторов и представленности книг на временной шкале с XVIII по середину XX века. Иначе говоря, в дополнение к датасету Google n-грамм,

очень большого, но непрозрачного по своим источникам, авторы исследования создают два «лабораторных» корпуса, параметры которых остаются максимально под их контролем. Совпадение результатов в большом датасете и в малых «дистиллированных» корпусах подтверждает валидность данных малых корпусов для поставленной задачи и открывает возможности к более глубокому исследованию факторов, которые могут влиять на снижение позитивно окрашенных эмоциональных слов. Авторы строят корреляционную модель, включающую в себя значения метаданных и доказывают, что именно дата издания книги является тем фактором, который наилучшим образом предсказывает динамику частотности эмоционально окрашенных слов. Авторы имеют возможность проверить и другие факторы, недоступные при работе с Google n-граммами, например, показывают отсутствие зависимости тренда от размера лексикона отдельной книги или от пола автора.

Еще одним примером развития методов работы с Google n-граммами является работа [Chalesworth, Caliskan, Banaji, 2022], посвященная исследованию исторических репрезентаций социальных групп в течение 200 лет. Авторы изучают, как менялись стереотипы 14 социальных групп, связанных с расой, национальностью, гендером, возрастом и внешностью. Основным методом стало исследование слов-ассоциатов (близких по значению), которые были получены с помощью анализа имбеддингов (т.е. общих контекстов) на материале 5-граммов датасета Google Books. Этот метод позволяет значительно расширить языковые списки исследуемых слов. Авторы используют не отдельные слова или группы слова для поиска, а список из 14 000 слов с позитивной или негативной окраской, выявляя те из них, которые имеют значимые контекстуальные связи с одной из групп. Таким образом, удается выявить наиболее устойчивые стереотипы, противопоставляющие одну группу другой (рис. 4.23).

Далее исследуется то, как меняются выделенные списки слов, связанные с определенной группой, с течением времени, а также в какой степени ассоциативные ряды каждой из противопоставленных групп пересекаются. Таким образом, удается проследить изменения или же стабильность социальных стереотипов в течение 200 лет.

Несмотря на то что работа [Michel et al., 2011] вводит в научный оборот термин культуромика именно как название метода анализа частотности слов книг Google (фактически статья является сопроводительной и иллюстрирующей возможности выложенного в открытый доступ ресурса Google n-грамм), развитие культуромики

Group A (vs. Group B)	Top 10 traits: Relative association	Valence
White (vs. Black)	Critical, polite, hostile, decisive, friendly, diplomatic, understanding, philosophical, able, belligerent	0.65
Black (vs. White)	Earthy, lonely, cruel, sensual, lifeless, deceitful, helpless, rebellious, meek, lazy	−1.18
Asian (vs. White)	Pompous, theatrical, verbal, superstitious, curious, traditional, melancholy, solemn, artificial, sensual	0.08
Irish (vs. White)	Passionate, pompous, melancholy, fanatical, headstrong, sly, grim, sarcastic, solemn, romantic	−0.45
Hispanic (vs. White)	Verbal, pompous, formal, solemn, abrupt, diplomatic, impetuous, traditional, evasive, lifeless	−0.63
Native American (vs. White)	Superstitious, rude, earthy, wholesome, spontaneous, artificial, kind, sensual, lonely, dependent	0.30
Men (vs. Women)	Able, competent, enterprising, honest, independent, brave, efficient, confident, practical, decisive	1.75
Women (vs. Men)	Charming, feminine, soft, romantic, modest, fair, lonely, gentle, tender, helpless	1.15
Old (vs. Young)	Traditional, pompous, solemn, humble, dignified, strict, grim, detached, diplomatic, possessive	−0.42
Young (vs. Old)	Vigorous, bright, hopeful, alert, fair, helpless, intellectual, thoughtless, patient, tender	0.97
Fat (vs. Thin)	Jolly, brave, honest, merry, generous, cheerful, wholesome, intelligent, compassionate, angry	2.06
Thin (vs. Fat)	Logical, theatrical, flexible, original, rigid, brilliant, superficial, detached, precise, artificial	0.43
Rich (vs. Poor)	Dominant, brilliant, dignified, conservative, decisive, respectable, diplomatic, independent, intellectual, artistic	1.45
Poor (vs. Rich)	Helpless, lonely, weak, lazy, dull, stupid, worried, ignorant, cold, careless	−1.85

Рис. 4.23. Для всех пар групп были выбраны два списка слов, каждый из которых содержит слова наиболее близкие по средней косинусной мере одной из групп относительно другой [Chalesworth, Caliskan, Banaji, 2022: 4]

как метода квантитативных исследований культурных и социальных трендов с помощью текстовых данных частотности слов и словосочетаний не связывается именно с датасетом Google n-грамм. Так, в статье [Silber-Varod et al., 2016] представлен обзор исследований, сделанных в русле идей культуромики (rise of culturomics [ibid: 84–85]) и посвященных анализу возникновения, развития и угасания технологических терминов и понятий. Следует отметить, что среди графиков, представленных в исходной статье, есть график, отражающий то, как в течение XIX–XX веков увеличивается скорость «забывания» нового изобретения [Michel et al., 2011: 179]. Термины, ввиду их однозначности и специфичности, оказываются очень хорошим объектом для запросов на больших языковых данных. Однако более глубокое развитие этого направления (культуромика истории науки, изобретений, институтов), как правило, происходит уже на специальных, пусть и менее обширных по временному размаху датасетах, которые обеспечивают более качественный отбор исходных данных, например, используются библиографические базы Web of Science [Chumtong & Kaldewey, 2017].

Еще одной альтернативой датасета Google n-gram может стать большой репрезентативный диахронический корпус, например национальный корпус языка. Так, в исследованиях по культуромике в работах [Бонч-Осмоловская, 2015, 2018] используются

данные Национального корпуса русского языка[1]. В статье [Бонч-Осмоловская, 2015] сравнивается динамика изменения частотности разных семантических классов прилагательных, характеризующих существительное *дорога* в разных аспектах его значения (маршрут, дорожное полотно, метафора пути, поездка и т.д.). В работе строятся наблюдения о том, какие из семантических классов ведут себя похожим образом, как меняется частотность вхождений семантических классов, а также их лексический состав с течением времени. Эти данные связываются с экстралингвистической реальностью. Таким образом, показывается, как изменения в общественной, социальной и культурной жизни находят отражение в изменениях языковой практики. В статье [Бонч-Осмоловская, 2018] рассматриваются конструкции, состоящие из прилагательного и названия десятилетия (*двадцатые, тридцатые, сороковые* и т.д.) и предпринимается попытка восстановить с помощью сравнения списков прилагательных и их частотностей мнемонический образ каждого десятилетия в «наивном» восприятии советской истории: то, каким образом тот или иной период времени остался в коллективной памяти, какой из периодов воспринимается как исторически значимый, а какой, наоборот, практически забыт. Особенностью этого исследования является то, что оно в большей степени качественное, а не квантитативное: частотность употребления того или иного прилагательного не так важна, как сам факт его употребления: характеристики десятилетия с помощью эпитета, будь то эмоционально окрашенное *гнилые восьмидесятые* или нейтральное *далекие пятидесятые*, неизбежно опираются на коллективные ассоциации, общие референции говорящего и слушающего к мнемоническому образу прошлого. Тем не менее рост частотности определенного прилагательного дает возможность увидеть, как авторская референция к комплексу воспоминаний трансформируется в повторяемый штамп. Именно это произошло со словосочетанием *лихие девяностые*. Несмотря на то что именно *девяностые* имеют самый богатый и семантически разнообразный арсенал прилагательных в датасете основного корпуса Национального корпуса русского языка, примерно с 2010-х годов практически все они вытесняются одним единственным прилагательным — *лихие*. Многократно повторенное, это выражение теряет свою функцию отсылки к коллективной памяти и начинает работать иным образом, являясь отсылкой к множеству его предшествующих

[1] www.ruscorpora.ru

употреблений и регламентируя основной способ отсылки к этому десятилетию (рис. 4.24).

Наиболее неожиданное развитие идей культуромики связано с областью экологии и охраны природы. Здесь возникло целое новое направление, название которого условно можно перевести как «культуромика охраны окружающей среды» (conservation culturomics). Дело в том, что природоохранная деятельность ставит своей задачей распространение информации о существующих угрозах и рисках окружающей среды и, таким образом, воздействие на общество. Задача conservation culturomics, таким образом, состоит в том, чтобы с помощью квантитативных методов выявить особенности общественного восприятия этих рисков, общественного интереса к природе и оценить, как пропаганда охраны окружающей среды в целом или экологические кампании, сфокусированные, например, на популяризации отдельного вымирающего вида, влияют на знание об угрозах, понимание важности охраны природы в обществе. С помощью conservation culturomics разрабатываются квантитативные метрики мониторинга окружающей среды и системы поддержки принятия решений в области охраны природы. Для этого могут быть использованы исследования на датасете Google n-граммов, отражающие длительные флуктуации (рис. 4.25), или же анализ данных последних лет, таких как статистика запросов в Google или сообщений в Твиттере (рис. 4.26). Принципиальным является то, что именно частотность релевантных слов и словосочетаний дает ключ

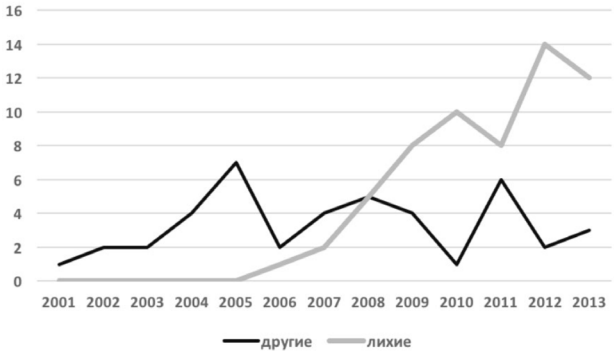


Рис. 4.24. Темная линия (все прилагательные, кроме *лихие*) вытесняется светлой линией (*лихие*) начиная с 2012 года, когда *лихие девяностые* становятся постоянно повторяемым речевым оборотом [Бонч-Осмоловская, 2018: 131]

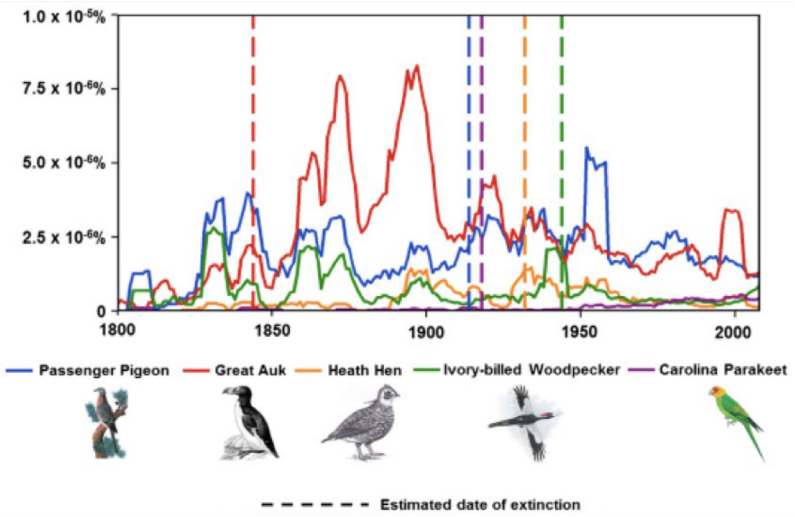


Рис. 4.25. Динамика частотности упоминаний четырех видов птиц в течение двух веков определенна с помощью Google n-граммов. Пунктиром обозначено время исчезновения каждого из видов [Ladle et al., 2016: 273]

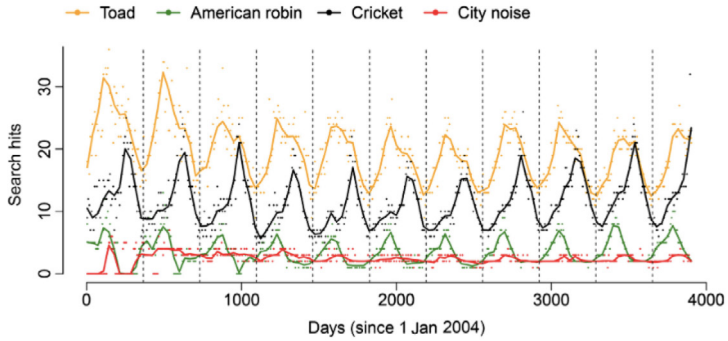


Рис. 4.26. [Ladle et al., 2016: 271] показывает «звуковую экологию» с помощью анализа запросов Google. Мы видим, что естественные шумы (*жаба, американский дрозд, сверчок*) имеют пики в летний период и падение в зимний, тогда как частота запроса *городской шум* практически не меняется в зависимости от времени года

к оценке значимости исследуемого явления в общественной жизни, а также, что даже более существенно, отражает динамику (рост или падение) его изменений. Подробный обзор проблем и подходов conservation culturomics представлен в работе [Ladle et al., 2016].

Проект культуромики является ответом на вызовы новой цифровой эпохи, сформулированные в статье [Halevy et al., 2009]: можно ли строить современные социальные и гуманитарные науки на фундаменте анализа больших данных, или иначе: можно ли исследовать культуру «дистантно» с помощью «сигналов» и «следов», отказавшись от медленного чтения и герменевтики текстов. Вне всякого сомнения, этот вектор научного развития является одним из ключевых в современных digital humanities и шире — например, в таких областях, как компьютерные исследования литературы, культурная аналитика, интернет-исследования в социологии и политологии. Но одновременно оказывается, что вызовом для социальных и гуманитарных наук является не столько сам факт использования больших данных, сколько их объяснительный потенциал. Достаточно ли простых неразмеченных данных, которых очень много? Действительно ли слово является однозначным и исчерпывающим сигналом для культуры, таким же, как последовательность ДНК в геноме? В работах московско-тартуской семиотической школы использовалось понятие «вторичных знаковых (моделирующих) систем» для описания того, как устроены различные формы культуры — от фольклора до философии. «Системы, в основе которых лежит натуральный язык и которые приобретают дополнительные сверхструктуры, создавая языки второй степени, удобно называть вторичными моделирующими системами. Искусство будет рассматриваться нами в ряду вторичных моделирующих систем» [Лотман, 1967]. Google Ngram Viewer дает нам информацию лишь об изменениях «первичных» частотностей естественного языка, но не о тех семиотических системах, в которых они участвуют. Как было показано, исследования, выполненные в духе культуромики, так или иначе достраивают «второй моделирующий этаж» — с помощью обогащения метаданных, с помощью жанровой системы, — если это исследования литературы, или же используя систему лексических конструкций, заданную в рамках определенной семиотической модели. В некотором смысле культуромика как проект, являясь прямым развитием манифеста [Halevy et al., 2009], одновременно опровергает его ключевую идею «данные вместо формул». Лучше всего об этом пишут Эйген и Мишель в своей книге «Неизведанная территория»,

посвященной культуромике как исследовательскому проекту: «Поскольку мы продолжаем накапливать необъясненные и недостаточно объясненные факты, появилось мнение, что причинно-следственная связь как основа научного познания рискует уступить свое место корреляции. Некоторым даже кажется, что дальнейшее развитие больших данных приведет к смерти теории. Однако с такой точкой зрения вряд ли можно согласиться. Мы можем отнести к подлинным триумфам современной науки такие теории, как теория общей относительности Эйнштейна или теория естественного отбора Дарвина, объясняющие причины сложных явлений с помощью небольшого набора основополагающих принципов. Если поиск таких теорий уйдет в прошлое, то мы рискуем потерять саму суть того, что называется наукой. Какой смысл делать миллионы открытий, если мы не можем объяснить сути ни одного из них? Это не значит, что мы должны отказываться от объяснений природы вещей. Это значит лишь, что мы должны изменить принципы своей работы» [Эрец Эйден & Мишель, 2016: 34].

Литература

Бонч-Осмоловская, А. (2015). Культуромика в Национальном корпусе русского языка, к постановке задачи: Три века русских дорог. *Труды Института русского языка им. В. В. Виноградова*, 6, 605–641.

Бонч-Осмоловская, А. (2018). Имена времени: Эпитеты десятилетий в Национальном корпусе русского языка как проекция культурной памяти. *Шаги / Steps*, 4(3–4), 115–146.

Даниэль, М.А., & Добрушина, Н.Р. (2016). *Два века в двадцати словах*. Национальный исследовательский университет «Высшая школа экономики». https://www.elibrary.ru/item.asp?id=29341745

Захаров В. П. & Масевич А. Ц. (2014). Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer. *Структурная и прикладная лингвистика*. 10, 303–327.

Лотман, Ю. (1967). Статьи по семиотике культуры и искусства (Серия «Мир искусств»). *Ученые записки Тартуского университета. Труды по знаковым системам.* 3(198), 130–145.

Эрец Эйден & Мишель, Ж.-Б. (2016). *Неизведанная территория: Как «большие данные» помогают раскрывать тайны прошлого и предсказывать будущее нашей культуры*. Издательство АСТ.

Charlesworth T. E. S., Caliskan A., Banaji M. R. (2022). Historical representations of social groups across 200 years of word embeddings from Google Books // Proceedings of the National Academy of Sciences of the United States of America. T. 119. № 28.

Chumtong, J., & Kaldewey, D. (2017). Beyond the google ngram viewer: Bibliographic databases and journal archives as tools for the quantitative analysis of scientific and meta-scientific concepts. *Forum Internationale Wissenschaft: Working Paper*, 08.

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Harmon, K. (2010, октябрь 17). *New Tool Tracks Culture through the Centuries via Google Books*. Scientific American. https://www.scientificamerican.com/article/google-books-culture/

Hilpert, M., & Gries, S. Th. (2016). Quantitative approaches to diachronic corpus linguistics. В *The Cambridge handbook of English historical linguistics* (Pp. 36–53).

Jha, A., & Kingsland, J. (2010, декабрь 17). *Culturomics and the new Google tool for tracking cultural trends*. The Guardian. http://www.theguardian.com/science/2010/dec/16/culturomics-google-tool-cultural-trends

K. Willems. (2013). "Culturomics" and the representation of the language of the Third Reich in digitized books. *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis*, *18.3*.

Ladle, R. J., Correia, R. A., Do, Y., Joo, G.-J., Malhado, A. C., Proulx, R., Roberge, J.-M., & Jepson, P. (2016). Conservation culturomics. *Frontiers in Ecology and the Environment*, *14*(5), 269–275. https://doi.org/10.1002/fee.1260

Michael, Robert and Karin Doerr (2002). *Nazi-Deutsch/Nazi German: An English Lexicon of the Language of the Third Reich.* CT: Greenwood.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Morin, O., & Acerbi, A. (2017). Birth of the cool: A two-centuries decline in emotional expression in Anglophone fiction. *Cognition and Emotion*, *31*(8), 1663–1675. https://doi.org/10.1080/02699931.2016.1260528

Ophir, S. (2016). Big data for the humanities using Google Ngrams: Discovering hidden patterns of conceptual trends. *First Monday*. https://doi.org/10.5210/fm.v21i7.5567

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLOS ONE*, *10*(10), e0137041. https://doi.org/10.1371/journal.pone.0137041

Ryan Heuser & Long Le-Hac. (2012). *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method. Panphlet 4. Stanford Literary Lab*. https://litlab.stanford.edu/assets/pdf/LiteraryLabPamphlet4.pdf

Shea, C. (2012, март 16). *The New Science of the Birth and Death of Words*.

Silber-Varod, V., Eshet-Alkalai, Y., & Geri, N. (2016). Culturomics: Reflections on the potential of big data discourse analysis methods for identifying research trends. *Online Journal of Applied Knowledge Management*, *4*(1), 82–98. https://doi.org/10.36965/OJAKM.2016.4(1)82–98

Zhang, S. (2015, октябрь 12). The Pitfalls of Using Google Ngram to Study Language. *Wired*. https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/

Глава 5

Базы данных: модели, структуры, связанные данные

(*Д. А. Гагарина*)

Цифровой поворот и структурирование информации в гуманитарных науках

Digital humanities — это разработка таких машиночитаемых моделей гуманитарных данных, гуманитарной информации, гуманитарных знаний, благодаря которым компьютер и программа становятся не просто средствами подсчета, а полноценными компаньонами исследователя. Цифровой поворот — это момент, когда получилось формализовать то, что только что казалось неформализуемым и неструктурируемым. Результаты и перспективы такой формализации интересны и пока в недостаточной степени проанализированы и отрефлексированы. Так, еще совсем недавно компьютер работал преимущественно с цифрами и текстами, в то время как рисунок был скорее набором пикселей. Сегодня технологии компьютерного зрения применяются к произведениям искусства и, например, отделяют сюжет картины от ее жанра и стиля и решают другие задачи истории искусства[1]. Корпусы и базы данных развиваются не только на основе текстов, но и других источников. Отсюда активно развивающимися направлениями становятся цифровая история искусства и цифровая история. В будущем мы, вероятно, сможем формализовать хранение и обработку эмоций, впечатлений, запахов и другой

[1] Oldman, Dominic, Diana Tanase, and Stephanie Santschi. (2019). "The Problem of Distance in Digital Art History: A ResearchSpace Case Study on Sequencing Hokusai Print Impressions to Form a Human Curated Network of Knowledge". International Journal for Digital Art History, no. 4 (December):5.29–5.45. https://doi.org/10.11588/dah.2019.4.72071; Pugh, Emily. (2020). "Art History Now: Technology, Information, and Practice". International Journal for Digital Art History, no. 4 (November):3.47–3.59. https://doi.org/10.11588/dah.2019.4.63448

информации, с которой имеют дело представители гуманитарных наук. Цифровизация идет неравномерно внутри гуманитарных наук в целом и внутри отдельных гуманитарных наук. То, в какой точке случается цифровой поворот в каждой конкретной проблемной области, в некотором смысле говорит о развитии этой области, о развитии цифровых технологий и о настроениях академического сообщества. Иногда область опережает технологии, иногда они с трудом адаптируются к друг другу.

Базы данных, являясь одной из традиционных и наиболее распространенных технологий, находят свое место в гуманитарных науках. Эта технология позволяет эффективно находить необходимую информацию, организовывать и систематизировать ее для более удобного доступа и анализа. Более того, с использованием современных технологий обработки данных, таких как машинное обучение и анализ больших данных, базы данных могут помочь в поиске новых связей между различными объектами гуманитарных наук и их свойствами, обеспечить новые возможности для исследований.

Базы данных, как один из базовых и при этом мощных инструментов формализации и структурирования информации, имеют особое значение в аспекте цифрового поворота. С одной стороны, их включенность в ту или иную область (в том числе гуманитарную) позволяет отследить масштабы цифровизации этой области, с другой — развитие самой технологии баз данных на протяжении десятилетий свидетельствует о состоянии инструментов и возможностей структурирования информации.

В этой главе речь пойдет о структурах и моделях данных и о том, как они организуются в базы данных и информационные системы, как происходит трансформация — цифровой переход — от источников (исторических, культурологических, лингвистических, литературных и других) к формальной структуре и далее — к базе данных и функциональному ресурсу на ее основе. Хотя моделирование, о котором будет идти речь в главе, имеет под собой математическую базу, мы не будем углубляться в нее, рекомендуя читателю обратиться к дополнительной литературе.

Специфика баз данных в гуманитарных науках

База данных — это организованная структурированная совокупность данных, которые хранятся и обрабатываются компьютером. Они организуются в соответствии со схемой (или моделью)

данных таким образом, чтобы с ними можно было эффективно работать, в частности хранить, управлять, извлекать данные в необходимом виде, обрабатывать их. Для создания баз данных используются СУБД — класс программ, позволяющих создать базу данных и манипулировать данными в ней. И базы данных, и СУБД отличаются значительным разнообразием, в первую очередь связанным с тем, какие модель и подход лежат в основе формализации информации и структурирования данных. Поэтому в этой главе мы рассматриваем одновременно и модели данных на уровне понятия и типов, и собственно базы данных и информационные системы.

Базы данных в гуманитарных науках важны не только в силу своей распространенности и эффективности, структурированная организация информации имеет особое значение для гуманитарных областей и для digital humanities — в некотором смысле концепция баз данных противопоставляется более привычному для гуманитарных областей текстовому представлению информации. Текстовое представление линейно, последовательно, в то время как база данных — это значительно более сложно связанные между собой объекты. «База данных становится центром творческого процесса в компьютерную эпоху», — пишет Лев Манович[1]. По мнению Чарльза Харви и Джона Пресса «проектирование и разработка баз данных играют центральную роль в трансформации методов исследований в истории»[2], это же во многом свойственно другим гуманитарным областям. Базы данных, по мнению Джоанны Друкер и соавторов, можно считать новой формой знаний, которая заменяет традиционные методы изучения истории и литературы, такие как повествование[3]. Это отчасти означает, что базы данных предоставляют более актуальную и персонализированную форму информации для исследователей и общества в целом.

В конце 90-х — начале 2000-х идут дискуссии и появляются обобщающие исследования по созданию и использованию баз данных в гуманитарных науках[4]. Необходимость разработки или адаптации

[1] Manovich L. Database as Symbolic Form // Convergence: The International Journal of Research into New Media Technologies. 1999. V. 5, is. 2.

[2] Harvey C., Press J. Databases in Historical Research. Theory, Methods and Application. London: Palgrave, 1996.

[3] Drucker J., Kim D., Salehian I., Bushong A. Introduction to Digital Humanities. Concepts, Methods, and Tutorials for Students and Instructors. Course Book. UCLA, 2013.

[4] Доорн П. Я и моя база данных: движение к концу направления «История и компьютинг» // Информационный бюллетень Ассоциации «История и ком-

методов и технологий баз данных для гуманитарных областей связана со спецификой исторических, филологических, культурологических данных. Они часто бывают неполными и противоречивыми, фрагментарными, неопределенными и неоднородными. Неполнота и фрагментарность связаны с целым комплексом причин — от традиций тех или иных научных областей и неравномерной их разработки до физической естественной или неестественной утраты исторических источников.

Ярким примером неоднородности является способ записи дат в исторических текстах, это может быть указание на конкретный день или только год, а иногда даже год неизвестен и идет указание на век или период. Классический тип данных «дата-время» в этом случае не подходит для задания признака, нужно придумывать специальные решения, например, с разделением признака на несколько. При этом в источниках может встречаться противоречивая хронологическая информация, когда одному событию приписываются разные даты, и для такого случая тоже необходимо иметь решение внутри базы. Это пример противоречивых данных, которые, конечно, могут быть не только у историков, и не только с датами, но и с другими объектами и их признаками.

Традиционные СУБД, которые разрабатывались скорее для задач технических и естественных наук, не всегда могут справиться со сложностью и многообразием гуманитарной информации. Ее специфика приводит не только к использованию и адаптации существующих СУБД, но и к разработке специализированных программных решений, ориентированных на гуманитарную информацию и гуманитарные задачи.

Один из наиболее ярких примеров специализированного программного обеспечения для решений гуманитарных задач — СУБД КЛИО, разработанная в 80-х годах Манфредом Таллером[1]. Эта система позволяла структурировать в виде базы исторические источники, для которых не подходили реляционные модели (которые

пьютер». 1995. № 13. С. 48–77; Harvey C., Press J. Databases in Historical Research. Theory, Methods and Application. London: Palgrave, 1996; Thaller M. Data bases v. critical editions // Historical Social Research. 1988. № 13 (3). Pp. 129–139; Корниенко С. И., Гагарина Д. А., Поврозник Н. Г. Исторические информационные системы: теория и практика. М.: Издательский дом НИУ ВШЭ, 2021; Denley P. Models, Sources and Users: Historical Database Design in the 1990s. // History and Computing. 1994. № 6. Pp. 33–44.

[1] Thaller M. Source Oriented Data Processing and Quantification: Distrustful Brothers // Historical Social Research. 1995. Supplement 29. Pp. 287–306.

мы подробнее рассмотрим ниже) и существующие на их основе в то время СУБД.

Другое значимое специализированное решение для структурирования гуманитарных источников — технология XML TEI[1], которая, по сути, стала лидирующей технологией семантической разметки текстовых источников в гуманитарных областях.

Модели и структуры данных, информации и знаний

От данных к мудрости

Данные, информация и знания — важные понятия не только для этой главы, но и для информатики в целом. Существуют разные концепции того, как логически выстраивается эта тройка понятий[2].

Центральным в тройке является понятие «информация». Собственно информатика — это наука об информации, при этом информатика появляется только в XX веке, в то время как информация, конечно, появляется намного раньше. Создание компьютеров как универсального средства обработки информации требует формализации этой самой информации, разработки того, как та или иная информация будет в памяти компьютера представлена. Десятки существующих определений информации не просто синонимичны, речь идет о принципиально разных подходах к пониманию информации.

Данные — самое простое понятие в рассматриваемой тройке, его можно определить, например, как представление фактов, понятий или инструкций в форме, приемлемой для общения, интерпретации, обработки человеком или с помощью автоматических средств. Данные и информация связаны. В соответствии с одним подходом, данные — это информация, представленная в форме, приемлемой для обработки, в том числе с помощью компьютера. Либо данные могут рассматриваться как зарегистрированная информация. Данные могут быть открытые, персональные, мы говорим о базах и банках

[1] The Text Encoding Initiative (TEI). https://tei-c.org/

[2] Например: Floridi L. Two Approaches to the Philosophy of Information. 2003. http://dx.doi.org/10.2139/ssrn.3853490; Zins C. Conceptual Approaches for Defining Data, Information, and Knowledge // Journal of the American Society for Information Science and Technology. 58 (4): 479–493, 2007. DOI: 10.1002/asi.20508; Rowley J. The wisdom hierarchy: representations of the DIKW hierarchy // Journal of Information and Communication Science. 33 (2): 163–180, 2007.

данных, о структурах, моделях и форматах данных. Данные могут быть текстовыми, пространственными, графическими и др.

Отношение между понятиями «информация» и «знания» не менее простые. Часто информация определяется через знания, то есть информация — некоторое структурированное и формализованное знание. Другой подход говорит о том, что в основании находится уровень данных, а информация добавляет к этому уровню контекст.

Знания — это более высокий уровень информации, который может быть получен из данных при помощи анализа, классификации, кластеризации и т.д., знание в некотором смысле добавляет к уровням данных и информации механизм использования. В этом подходе есть еще и четвертое понятие — мудрость, которая добавляет условия использования знаний. Сам подход называется DIKW — как аббревиатура от data, information, knowledge, wisdom[1].

В контексте этой главы важно разобраться, как информацию, относящуюся к гуманитарному полю, представить в цифровом виде, то есть как перейти от информации, получаемой при изучении источника, к данным, которыми этот источник будет представлен в памяти компьютера.

Модели или структуры данных — это способы организации информации, которые используются для ее хранения, передачи и обработки. Существуют разные классификаторы и типологии моделей данных в зависимости от решаемых задач, далее — внутри каждой классификации каждая модель или структура данных опять же предназначена для решения определенных задач и имеет свои преимущества и ограничения по возможностям хранения, передачи или обработки. Модели и структуры данных являются основой для работы с информацией, а анализ и извлечение знаний — целью этой работы. В гуманитарных науках, где источники часто имеют неструктурированный характер, их формализация и анализ становятся нетривиальными задачами.

[1] Rowley J. The wisdom hierarchy: representations of the DIKW hierarchy // Journal of Information and Communication Science. 33 (2): 163–180, 2007.

Типы данных и семантический разрыв

История компьютеров и технологий связана с развитием способности компьютеров понимать и обрабатывать различные типы информации. В начале развития электронно-вычислительных машин ставилась задача обработки числовой информации, которую можно представить в двоичной системе, что удобно для физической реализации в компьютере.

Следующим типом информации, возможным для хранения в памяти компьютера, стала текстовая информация, которую можно закодировать числовыми кодами символов, но этого оказалось недостаточно, так как кодирование отдельных символов имеет огромный семантический разрыв с теми задачами, которые ставятся при обработке текстов.

Перейдем к кодированию графики — еще одному типу информации, ставшему доступным для компьютера. Рисунок можно формализовать, разбив его на пиксели с координатами и цветом, — это называется пиксельной графикой. Помимо пиксельной, существует векторная графика, элементами которой являются не точки-пиксели, а линии.

Дальше была аудиоинформация, затем видео, потом все более сложные виды информации, включая, например, эмоции, которые тоже можно формализовать и задать математически.

Одной из задач в цифровых гуманитарных исследованиях является использование компьютеров и программ для анализа гуманитарной информации. Однако восприятие текста-источника исследователем, будь то историк или культуролог, значительно отличается от представления текста в виде набора символов. Аналогично набор пикселей на картине не является тем, что видит искусствовед. В результате через набор символов или набор пикселей задается настолько слабая и примитивная модель, что это не позволяет компьютеру проводить с источником аналитические исследования, кроме подсчета количества букв или пикселей. Это демонстрирует семантический разрыв. Преодоление этого разрыва — это поиск или разработка для интересующего нас источника такой формальной модели, которая максимально точно передаст его восприятие человеком. Тогда мы сможем делегировать на компьютер некоторые функции человека по анализу источника, а затем и выйти за пределы этих функций. Преодоление семантического разрыва требует глубокого понимания

предметной области и сочетания цифровых методов исследования с традиционными методами.

Еще одно важное понятие в связи с типами данных и представлением информации — машиночитаемые данные. Машиночитаемость (machine-readable) относится к возможности декодировать информацию из ее формы представления и интерпретировать ее программными средствами. Аналогично существуют человекочитаемые данные. Например, если имеется фотография текстового документа, то его можно прочитать глазами, но для компьютера это будет просто набором пикселей, которые невозможно интерпретировать как текст. Для того чтобы фотография текста стала текстом в машинном понимании, необходимо ее распознать, используя технологию OCR (Optical Character Recognition). Документы в форматах, таких как Word или PDF, являются примерами форматов, которые читаемы как человеком, так и машиной. Тем не менее более сложные модели текстовых документов, которые предполагают глубокую семантическую разметку, могут быть нечитаемы для человека, но при этом могут позволить автоматические манипуляции.

Структуры данных и их использование в гуманитарных областях

Объекты, относящиеся к данным разного типа, объединяются в структуры (схемы, модели). Эти структуры можно классифицировать в соответствии с типами связей между объектами. К типичным структурам относятся линейная, иерархическая, сетевая и реляционная. Три последние из них могут быть основой базы данных. Собственно исторически так и было: сначала появились иерархические базы данных, затем сетевые, а потом ограничения и возможности сетевых и иерархических моделей были учтены в реляционной модели.

Линейная модель

Линейная модель — самая простая, ее элементы последовательно меняются друг за другом. У каждого элемента, кроме первого и последнего, есть один элемент на входе и один на выходе. Примерами линейной модели являются любые последовательности

событий, списки людей, документов. Интересно, что и текст в некотором смысле линеен, так как в нем слова следуют за друг за другом, предложения тоже. И если слой слов и предложений рассматривать по отдельности, то каждый из них будет вполне линейной моделью.

Вариантом линейной модели может быть двумерная таблица — структура, в которой информация хранится в виде линейно упорядоченных записей, называемых также кортежами или строками. В этой модели данные хранятся в таблице, которая состоит из столбцов и строк. Каждый столбец представляет собой тип данных, такой как число, дата или текст, а каждая строка представляет собой уникальную запись. В этой модели данных отсутствует связь между таблицами, что может привести к повторению информации в различных строках, ухудшению производительности и сложности обработки данных, однако модель часто используется в небольших приложениях или для хранения небольших объемов данных. Ее недостаток — неэффективность при работе с разнородными и многомерными данными. Примеры: исторические архивы могут использовать линейную модель данных для хранения информации о документах, включая название документа, дату создания, автора и т.д.; линейная модель данных также может применяться для хранения информации о литературных произведениях, их авторах, жанрах и т.д.

Иерархическая модель

Иерархическая модель данных — это структура базы данных, в которой данные представляются в виде иерархии, подобной дереву, где каждый узел имеет одного родителя, кроме корневого узла, у которого нет родителя. В этой модели данных каждый узел может иметь несколько дочерних узлов, но только одного родителя. Каждый узел представляет собой отдельную сущность или запись, а связи между узлами задаются в виде ссылок на родительские узлы. Иерархическая модель достаточно проста для понимания, и на ее основе организуются самые разные данные, однако ее значимым ограничением является жесткая структура. Она не позволяет задать горизонтальные связи или сделать для элемента два «родителя».

Иерархическая модель данных была разработана в 1960-х годах и широко применялась в ранних базах данных, таких как IMS

(Information Management System) от IBM[1]. С появлением сетевых, реляционных и других более гибких моделей данных использование иерархической модели данных снизилось. Тем не менее в гуманитарных науках эта модель данных все еще может быть полезна для представления данных с явной иерархической структурой, таких как генеалогические деревья, библиотечные каталоги, структуры организаций и сообществ и т.д.

Иерархическим может быть и текст в том смысле, что он делится на главы, глава делится на параграфы, параграф на абзацы и так далее. Именно такая иерархическая модель текста лежит в основе разметки XML TEI.

Сетевая модель

Сетевая модель данных представляет собой структуру, в которой каждый элемент данных может иметь несколько связей с другими элементами, образуя сложную сеть связей. Эта модель представляет собой расширение жесткой иерархической модели данных, позволяя описывать сложные связи между элементами данных, которые не всегда могут быть представлены в виде иерархической структуры. Впрочем, эта гибкость является во многом ограничением модели, запутанные связи усложняют процесс обработки данных.

Сетевая модель позволяет описывать сложные и многомерные взаимосвязи между данными, что делает их полезными и эффективными для гуманитарных наук, она используется для моделирования социальных сетей, культурных связей, генеалогических деревьев, семантических сетей и прочих форм связей между людьми, текстами, идеями, организациями и другими объектами. Например, сетевая модель может быть использована для анализа социальных сетей в сообществах, где связи между людьми могут быть сложными и многоуровневыми; для анализа лингвистических структур, таких как семантические сети, которые отображают связи между словами и понятиями.

Одним из преимуществ сетевых моделей данных является возможность учитывать многомерные связи между узлами и связями. Это позволяет проводить более глубокий и точный анализ социальных и культурных явлений. Кроме того, сетевые модели

[1] Blackman K. R. Technical note: IMS celebrates thirty years as an IBM product // IBM Systems Journal, vol. 37, no. 4, pp. 596–603, 1998. DOI: 10.1147/sj.374.0596.

могут быть использованы для создания графических визуализаций данных, что облегчает их восприятие и понимание.

Однако сетевые модели данных могут быть сложны в реализации, они менее универсальны, чем реляционные и иерархические базы. Сетевые модели могут быть реализованы с помощью специализированных графовых баз данных, которые предоставляют инструменты для анализа и манипулирования сетевыми структурами. Эти базы данных позволяют быстро и эффективно хранить, искать и анализировать связанные данные.

Реляционная модель

Реляционная модель, предложенная Эдгаром Коддом в 1970 году[1], преодолевает ограничения сетевой и иерархической моделей, сохраняя при этом их преимущества и возможности. Это один из наиболее распространенных и используемых типов моделей баз данных, в том числе в гуманитарных науках. Реляционные модели основаны на концепции таблиц и отношений между ними, что делает их удобными для хранения и управления большим объемом данных, которые могут быть связаны между собой.

Реляционные модели позволяют организовывать данные в связанные между собой таблицы, что делает их более удобными для анализа и поиска. В таблицах можно использовать различные типы данных, такие как числа, тексты и даты, при этом реляционные модели данных позволяют связывать данные из разных таблиц, что дает возможность реализовывать структуры сложного уровня.

Реляционная модель получает активное развитие на рынке баз данных с 80-х годов.

Конечно, реляционная модель не лишена недостатков. Хотя ее многолетнее применение для гуманитарных областей доказало эффективность, оно сопряжено с некоторыми трудностями. Эти трудности определяются требованиями реляционной модели, согласно которым необходимо строго задать схему данных уже на начальном этапе. Для гуманитарных предметных областей это иногда сложно или даже невозможно, так как при изучении предметной области могут появляться новые объекты или меняться связи между существующими элементами и объектами.

[1] Codd E.F. A Relational Model of Data for Large Shared Data Banks // Communications of the ACM. 13 (6): 377–387, 1970. Doi: 10.1145/362384.362685.

Некоторые ограничения реляционной модели преодолеваются в объектно-реляционных, объектно-ориентированных и NoSQL базах данных.

NoSQL

Некоторые ограничения реляционной модели позволяют преодолеть NoSQL (Not only SQL) базы данных. Это класс решений для баз данных, которые не требуют жесткой схемы данных на входе, соответственно такие базы данных более гибки и масштабируемы. NoSQL отличается от реляционных моделей тем, что не использует традиционные таблицу, строки и колонки для хранения данных. Вместо этого NoSQL хранят данные в, ключ-значение, столбцы, и т.д. Преимуществом NoSQL баз данных является возможность работать с большими объемами неструктурированных данных, обеспечивая при этом хорошую производительность.

Технология получает интересное применение для гуманитарных областей, NoSQL базы данных могут использоваться для работы с большим объемом неструктурированных данных, таких как тексты, фотографии, видео, звуковые файлы и др. Например, для хранения и обработки большого количества литературных произведений или аудиозаписей можно использовать документо-ориентированные базы данных. Другой пример применения NoSQL баз данных в гуманитарных науках — работа с графовыми данными. Графовые базы данных используются для хранения и обработки данных, связанных с сущностями и их отношениями друг с другом. Такие базы данных могут использоваться для анализа социальных сетей, исследования текстов и других приложений.

Ограничения NoSQL баз данных в гуманитарных науках обусловлены слабой структурированностью данных, что может быть менее удобочитаемым и более трудно обрабатываемым, чем данные в реляционных базах данных. При этом NoSQL базы данных могут быть менее подходящими для хранения и обработки таких структурированных данных, как таблицы и связи между ними, что делает их менее универсальными по сравнению с реляционными моделями.

Примеры NoSQL баз данных, которые могут использоваться в гуманитарных областях:

– MongoDB — документо-ориентированная база данных, которая хранит данные в формате BSON (Binary JSON), MongoDB позволяет гибко организовывать данные и сложные структуры;

– Cassandra — распределенная колоночная база данных для обработки больших объемов данных;

– Neo4j — графовая база данных, ориентированная на хранение и обработку связанных данных и моделирование социальных сетей и других подобных задач;

– Couchbase — распределенная многомодельная NoSQL база данных, которая поддерживает гибкие модели данных, такие как документо-ориентированные, колоночные, ключ-значение и графовые модели, используются структурированные и неструктурированные данные;

– Firebase — облачная платформа от Google, предоставляющая инструменты для разработки приложений и хранения данных, в том числе NoSQL базу данных.

Проектирование гуманитарно ориентированной базы данных

Переходя к вопросам проектирования баз данных необходимо определить типичные модели гуманитарно ориентированных баз данных с учетом многообразия их назначения и контента. Для решения этой задачи сначала остановимся на типичных сущностях и возможных связях между ними.

Типичные сущности гуманитарно ориентированной базы данных

Обратимся к тому, на основе каких объектов (сущностей) создаются базы данных в гуманитарных областях. Природа этих сущностей весьма разнообразна, ими могут быть материальные объекты — архитектурные и другие сооружения, орудия труда, оружие, предметы домашнего обихода, одежда и другие памятники материальной культуры. Внутри одной базы данных может быть сразу несколько сущностей, связанных с материальными памятниками, или одна сущность, которая их объединяет.

База данных может включать информацию о людях, тогда основной ее сущностью становится человек, а в базе хранится демографическая, социокультурная и профессиональная информация об изучаемой группе людей. Такие базы называются просопографическими,

на их основе можно формировать коллективные портреты, выделять и изучать сообщества. Официальные и неофициальные сообщества, организации и группы людей тоже могут быть сущностью базы данных.

Другим и, пожалуй, наиболее типичным объектом баз данных являются источники и документы в самом широком смысле — как текстовые, так и аудиовизуальные и любые другие. В этом случае в базе могут храниться метаданные, полные тексты, цифровые копии источника.

То, как выделяются сущности базы данных, связано как со спецификой предметной области, так и задачами, для решения которых создается база данных. Внутри одной базы могут быть сущности разной природы, например, персоны и связанные с ними документы и события.

В наших ранних работах выделены и подробно рассмотрены типичные сущности исторических баз данных[1], среди которых:

- Персона — историческая личность или любой человек, среди типичных атрибутов биографические, социокультурные, профессиональные атрибуты: ФИО, дата рождения, дата смерти, пол, социальная принадлежность, профессия, образование, семейное положение и т.д.
- Источник — исторический источник любого типа, среди типичных атрибутов традиционные источниковедческие признаки, тематические, хронологические, археографические атрибуты.
- Публикация — научная, справочно-энциклопедическая или иного вида публикация, массив которых используется как содержательная или структурная основа базы, среди атрибутов библиографические, тематические.
- Организация (сообщество) — структурная единица, соответствующая организации или институции любого типа — общественная, государственная, политическая, научная, коммерческая, некоммерческая, юридическая и т.д., атрибуты определяются типом организации, ее структурой и функционированием.
- Событие — происшествие, явление или иная деятельность как факт государственной, общественной или личной

[1] Корниенко С. И., Гагарина Д. А., Поврозник Н. Г. Исторические информационные системы: теория и практика. М.: Издательский дом НИУ ВШЭ, 2021; Гагарина Д. А. Моделирование в истории: подходы, методы, исследования // Вестник Пермского университета. Серия: Математика. Механика. Информатика. 2009. № 7.

жизни, описываемое в системе, характеризуется в первую очередь пространственно-временными атрибутами.

Связи между объектами

Разработка модели данных предполагает наряду с выделением сущностей определение связей между ними. Такие связи тоже можно классифицировать. Так, существует три типа связей между сущностями внутри модели данных: один к одному, один ко многим, многие ко многим.

Один к одному (one-to-one): один элемент первой сущности связан не более, чем с одним элементом второй сущности, то есть одна запись в таблице A может быть связана только с одной записью в таблице B, и наоборот.

Например, каждый студент может иметь только один номер студенческого билета, а каждый номер студенческого билета может относиться только к одному студенту.

Другой пример: пусть у нас есть база музеев и их директоров на некоторый момент времени. Тогда каждому музею, то есть каждому элементу сущности Музей, в базе соответствует один элемент сущности Директор, и наоборот. Пусть в нашей базе музеев есть сущность, которая хранит странички музеев на некоторой платформе, например, русскоязычной Википедии. Тогда для каждого музея есть только одна страница в Википедии, и наоборот. Тоже связь один к одному.

Связь один к одному — по сути это разделение одной сущности на две, так в примере выше директора музея можно было бы хранить всего лишь как поле в таблице музеев, а не как отдельную сущность. В этой же таблице можно хранить ссылку на страницу Википедии. То есть все три сущности в примере можно без потерь соединить в одну. В некоторых случаях разделение сущности на две или три со связями один к одному может оправдываться природой сущностей или целями более эффективной последующей обработки данных.

Один ко многим (one-to-many): одному элементу первой сущности может соответствовать несколько элементов второй сущности, то есть одна запись в таблице A может быть связана с несколькими записями в таблице B. При этом обратное неверно, то есть одному элементу второй сущности (таблицы B) соответствует только один элемент первой сущности (таблицы A). Связь один ко многим — самая популярная для реляционных моделей. Аналогична

рассмотренной связь многие к одному — просто меняем первую и вторую сущность местами.

Например, каждый проект может иметь несколько задач, но каждая задача относится только к одному проекту.

И вернемся к примеру с музеями. Пусть в нашей базе хранится информация о музеях и картинах. Связь между объектами Музей и Картина — один ко многим, потому что одному музею в базе соответствует несколько картин, при этом каждой картине соответствует только один музей — тот, в котором она хранится.

Многие ко многим (many-to-many): каждый элемент первой сущности (таблицы A) может быть связан с множеством элементов второй сущности (таблицы B), и наоборот. Такие связи реализуются через дополнительную таблицу-связку (таблицку ассоциации), которая содержит пары ключей из обеих таблиц, то есть связь многие ко многим реализуется через две связи один ко многим.

Например, каждый студент может изучать несколько предметов, и каждый предмет может изучаться несколькими студентами. Таблица-связка будет содержать пары идентификаторов студентов и предметов.

Усложним модель в примере с музеями. Пусть наша база хранит еще и художников. Каждому музею в базе может соответствовать несколько художников, произведения которых представлены в этом музее. А каждому художнику соответствует несколько музеев, где хранятся его произведения.

Этапы проектирования базы данных

Разработка баз данных, в том числе гуманитарных, включает несколько этапов проектирования и жизненного цикла:

- Анализ требований: на этом этапе определяются цели и задачи, которые должна решать база данных, и требования к ее структуре и функциональности.
- Проектирование концептуальной модели: на этом этапе выделяются сущности будущей базы и связи между ними, на основе которых будет создана база данных.
- Проектирование логической модели: на этом этапе концептуальная модель преобразуется в логическую модель, в которой определяются отношения между сущностями, а также атрибуты

и ограничения. То есть на этом этапе набор выявленных ранее сущностей и связей необходимо представить в виде одной из существующих моделей — например, реляционной.

- Проектирование физической модели: на этом этапе логическая модель преобразуется в физическую модель, которая предполагает реализацию модели в выбранной СУБД и включает определение таблиц, индексов, полей и других элементов, необходимых для создания базы данных.
- Разработка приложения (информационной системы) для управления базой данных: после проектирования физической модели можно приступать к разработке приложения, которое будет использоваться при работе с базой данных.
- Ввод данных.
- Эксплуатация базы данных.

Для концептуального и логического проектирования существуют специализированные программные средства, хотя эти задачи могут успешно решаться «вручную на бумаге».

Подходы к созданию гуманитарно ориентированных баз данных

Помимо различий в типах собственно структур баз данных, можно выделить подходы к тому, на основе чего строятся эти структуры, как выделяются сущности и определяются связи. Значимая работа и дискуссии в этом направлении велись в 90-е годы прошлого века по отношению к историческим базам данных, однако во многом их результаты можно отнести и к другим гуманитарным областям. Возможность такого переноса связана с широтой и разнообразием предметов изучения исторической науки.

Важным шагом в понимании проблем создания исторических (историко-ориентированных) баз данных стали работы Манфреда Таллера и Питера Доорна[1], было выделено два основных подхода:

[1] Доорн П. Я и моя база данных: движение к концу направления «История и компьютинг» // Информационный бюллетень Ассоциации «История и компьютер». 1995. № 13. С. 48–77; Таллер М. Что такое «источнико-ориентированная база данных»; что такое «историческая информатика»? // История и компьютер: новые технологии в исторических исследованиях и образовании / под ред. Л. И. Бородкина и В. Левермана. Goettingen, 1993; Корниенко С. И., Гагарина Д. А.,

проблемно-ориентированный и источнико-ориентированный. В дополнение к этим подходам используется методо-ориентированный, а наиболее частым вариантом является смешанный. Остановимся на них подробнее.

Все подходы обладают рядом преимуществ и ограничений, их понимание важно для правильного выбора исходных оснований для проектирования системы. Хотя на практике часто используется смешанный подход, модули такой системы обычно строятся на основе одного из двух указанных подходов.

Источнико-ориентированный подход к созданию баз данных в гуманитарных науках предполагает создание структуры базы на основе структуры источника или массива источников, чаще всего однотипных.

Источнико-ориентированный подход сохраняет источник с его структурой, содержанием и внешним видом, что позволяет эффективно сохранить и расширить доступ к историческим источникам. Одним из главных преимуществ этого подхода является возможность многократного использования созданной базы данных для исследований или решения других задач.

Этот подход может быть неэффективным для уникальных источников или коллекций источников разнообразной формы и структуры из-за невозможности или сложности создания унифицированной структуры, подходящей ко всем источникам в изучаемой коллекции.

Проблемно-ориентированный подход является альтернативой источнико-ориентированному подходу и предполагает ориентацию на конкретную проблему или вопрос, который исследователь хочет решить с помощью базы данных. Этот подход подразумевает, что сначала формулируются гипотеза, конкретный вопрос или группа вопросов в пределах предметной области, на основе этого и анализа предметной области строится ее модель. Проблемно-ориентированный подход может использоваться в случаях, когда на начало исследования и разработки базы данных нет четкого понимания о том, какие источники нужны для решения определенной проблемы, когда источники разнообразны по структуре или слабо структурированы. При проблемно-ориентированном подходе для заполнения базы данных подбираются источники, содержащие

Поврозник Н.Г. Исторические информационные системы: теория и практика. М.: Издательский дом НИУ ВШЭ, 2021.

информацию по выбранным вопросам, но информация, которая имеется в источниках по другим вопросам, опускается. При этом используются разные источники, что является менее жестким подходом с точки зрения требований к источникам.

Этот подход позволяет более эффективно работать с неструктурированной информацией и решать конкретные проблемы в гуманитарных науках. Основным ограничением является то, что происходит частичная или полная потеря источника для повторного использования; при этом подходе можно столкнуться с проблемой противоречивости данных в источниках.

Методо-ориентированный подход к созданию баз данных фокусируется на тех или иных методах исследований, логика и идея этих методов становится основой структурирования данных. Ярким примером являются просопографические базы данных, отсылающие к просопографическому методу построения коллективных портретов.

На практике большинство баз данных создаются на основе смешанного подхода, использующего и источнико-, и проблемно-ориентированные модули. Обычно общая схема и функциональные модули разрабатываются при использовании проблемно-ориентированного подхода, а модули, содержащие источники, — источнико-ориентированного. Это позволяет сочетать преимущества обоих методов, сохранять источники для последующего использования и эффективно решать конкретные задачи на основе созданной системы.

Связанные данные и гуманитарные науки

Следующим уровнем моделирования данных, информации и знаний после создания на их основе структур и баз данных является концепция связанных данных (Linked Data). В случае Linked Data связываются различные источники информации и различные базы, и можно использовать их для создания более широкой и глубокой картины исследуемых явлений. Сам термин и концепция были предложены для веб-публикации информации, однако их применение выходит за пределы веб-среды. Для реализации Linked Data используются такие форматы и стандарты, как RDF (Resource Description Framework), OWL (Web Ontology Language) и др.

Связанные данные играют важную роль в гуманитарных науках, занимающихся исследованием человеческой культуры и истории, информация о которых распределена по различным источникам и представлена в разных форматах. Это затрудняет объединение информации и использование ее для создания новых знаний и выводов.

Кроме того, связанные данные позволяют исследователям лучше организовывать и представлять свои собственные данные. Используя стандарты и форматы связанных данных, можно создавать онтологии и семантические модели для стандартизации и организации данных.

Связанные данные представляют собой мощный инструмент для гуманитарных исследований, который помогает улучшить доступность, организацию и интерпретацию данных, однако в настоящее время их потенциал недостаточно раскрыт на практике и не получил должного распространения.

Глава 6

Компьютерный анализ текста

(*Б. В. Орехов*)

Цифровое исследование в гуманитарной науке оказывается возможным там, где удачным образом сходятся три компонента: данные, методы и исследовательский вопрос.

Наиболее удобным источником данных для количественных исследований является текст. В свернутом виде он содержит информацию, релевантную для историков, культурологов, лингвистов и литературоведов. Есть случаи, когда зафиксированная в текстовой форме информация оказалась востребована и представителями естественных наук[1].

Текстовые данные занимают не так много места, как мультимедийные файлы, они проще для оцифровки и обильно представлены в современном интернете.

Методы для машинного анализа текста начали разрабатываться задолго до современного компьютерного бума инженерной областью, которая называется компьютерной лингвистикой. В большинстве своем эти методы основываются на идее значимости частоты, с которой те или иные языковые единицы (например, слова) встречаются в тексте. Идея частотности делает осмысленной операцию подсчета, на которой основывается цифровой взгляд на гуманитарный материал.

С формированием современной ситуации, подразумевающей доступность больших вычислительных мощностей и свободно распространяемые крупные текстовые архивы, разработка методов пошла быстрее. Многие хорошо зарекомендовавшие себя технологии

[1] Neuhäuser R. et al. Colour evolution of Betelgeuse and Antares over two millennia, derived from historical records, as a new constraint on mass and age // Monthly Notices of the Royal Astronomical Society. 2022. Т. 516. № 1. С. 693–719.

оформлены в виде функций открытых программных библиотек, прежде всего, для языка Python.

Исследовательские вопросы вырастают из конкретной предметной области, из ее традиций и наработок. Часто это могут быть попытки уточнить какие-то положения, сформулированные без применения количественных методов, добавить дигитальные аргументы в теоретическую дискуссию, проверить релевантность ранее сделанных выводов на большом объеме данных.

Подготовка текстовых данных для машинного анализа

Работа с текстовыми данными предполагает, что мы подготовим их до применения компьютерных аналитических инструментов. Это касается не только текста, но и любых данных, у каждого типа данных есть своя специфика, отражающаяся на их предварительной обработке. Для изображений иногда требуется представить имеющуюся в нашем распоряжении коллекцию так, чтобы все они были одинакового размера и в одинаковом формате (например, содержали одинаковое число каналов для передачи цвета). Иногда нужно «убрать» все цвета и сделать изображение черно-белым. Такие же операции унификации данных (или, как еще говорят, препроцессинга) и их избавления от «шума» существуют для текста.

У всех операций, которые входят в препроцессинг, есть свой исследовательский смысл, поэтому их не следует рассматривать как чисто механические. Их выполнение полностью зависит от конечной исследовательской задачи и тех методов, которые мы избираем для ее решения. Например, лемматизация (см. ниже) уместна перед тематическим моделированием, но является избыточной для задач стилеметрии.

Данные, которые мы используем в исследованиях, существуют в своего рода «дикой природе». Это означает, что они отягощены контекстом своего бытования. Есть обстоятельства, при которых этот контекст нам не важен или даже мешает, и его следует отделить от данных или просто удалить до анализа.

Так, например, большое количество текстовых данных содержится в интернете. Развитие Всемирной сети вообще сильно повлияло на развитие науки о данных: до широкого распространения интернета

основное (если говорить образно) «топливо» науки найти было не так просто. Появлением многих современных технологий обработки текста мы тоже обязаны тому, что интернет стал огромным, и в нем хранится много текстовой информации. Такие технологии основаны на статистике распределения слов и поэтому чувствительны к объемам информации, с которыми они работают. Если данных мало, статистика дает сбои, показывает не тенденции, а шум, случайные значения. Если данных много, статистика с помощью измерений позволяет устанавливать закономерности. Поэтому, если бы у нас не было интернета в качестве почти неисчерпаемого источника текстов, мы бы не смогли обучать современные нейросетевые модели, полностью основанные на статистической индукции.

Но тексты берутся не только из интернета. Для цифровых гуманитарных исследований одним из источников текстовых данных является оцифровка документов. Документами в терминологии компьютерной лингвистики и информационного поиска называются и книги, и рукописи, и плакаты, то есть любые объекты, на которых может быть что-то написано или изображено.

Существуют очевидные и не слишком очевидные особенности предобработки текстов, с которыми исследователю приходится сталкиваться перед тем, как перейти к анализу.

Среди очевидного — ошибки распознавания символов, которые нужно исправлять. Оптическое распознавание символов, по-английски сокращенно называется OCR (optical character recognition), — это то, с чем сталкивается любой специалист, работающий с оцифровкой документов.

Менее очевидно, что в том же интернете тексты размещены на HTML-странице, а страницы содержат код, который помогает их отображению.

Например, вот такое представление имеет HTML-код вокруг содержательного текста на странице сайта rvb.ru (Русская виртуальная библиотека):

```
<hr class="color-19">

<h1>ПОСЛАНИЕ К ЖУКОВСКОМУ В ДЕРЕВНЮ</h1>
<div class="versusia6">
<span class="line" id="L1">Итак, мой милый друг, оставя
скучный свет</span><br>
```

```
<span class="line" id="L2">И в поле уклонясь от шума
и сует,</span><br>
<span class="line" id="L3">В деревне ты живешь, спокойный
друг природы,</span><br>
<span class="line" id="L4">Среди кудрявых рощ, под сению
свободы!</span><br>
<span class="line" id="L5">И жизнь твоя течет, как свет-
лый ручеек,</span><br>
<span class="line" id="L6">Бегущий по лугам, как легкий
ветерок,</span><br>
<span class="line" id="L7">Играющий в полях с душистыми
цветами</span><br>
<span class="line" id="L8">Или в тени древес пастушки
с волосами.</span><br>
```

Этот код тоже представляет собой текст, но для исследования он не важен. Терминологически такой код называется «обвязкой». С помощью кода текст на веб-странице перемежается с рекламой и видеоплеерами.

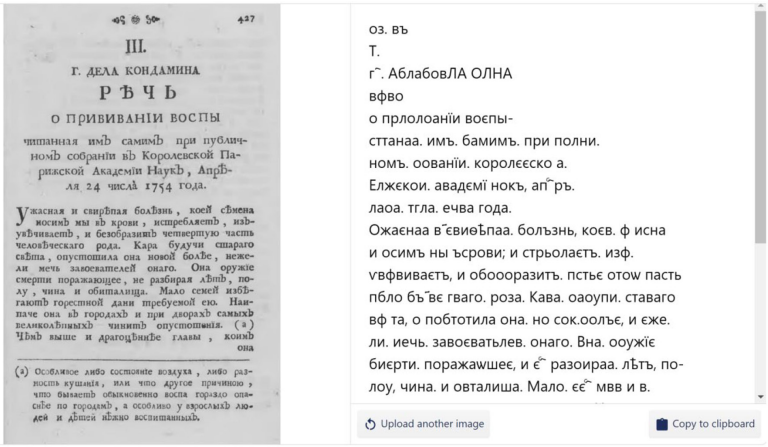
```
<span class="line" id="L27">То лирою своей Климену вос-
хищаешь,</span><br>
<span class="line" id="L28">То быстро на коне несешься
по полям,</span></div>
<div class="page" id="pg49">49</div>

<div id="yandex_rtb_R-A"></div>
<script type="81-text/javascript">window.yaContextCb.
push(()=>{
Ya. Context.AdvManager.render({
renderTo: 'yandex_rtb_R-A',
blockId: 'R-A-1281369-4'
})
})</script>
</div>

<div class="versusia6">
<span class="line" id="L29"><i>Как шумный ветр пустынь;</
i> то ходишь по утрам</span><br>
```

```
<span class="line" id="L30">С собакой и ружьем — и с пти-
цами воюешь;</span><br>
<span class="line" id="L31">То, сидя на холме, прелестный
вид рисуешь!</span><br>
```

При сканировании и распознавании печатных источников исследователь обязательно сталкивается с ошибками распознавания символов. Несмотря на то что есть типы текстов, в которых такие ошибки сводятся к минимуму (например, это современные стандартизованные книги), много проблем остается при оцифровке старых печатных источников. Например, программы для OCR плохо понимают дореволюционную орфографию и использовавшиеся в XVIII и XIX веках шрифты.



427

III.

Г. ДЕЛА КОНДАМИНА

РѢЧЬ

О ПРИВИВАНІИ ВОСПЫ

читанная имѣ самимѣ при публичномѣ собраніи вѣ Королевской Парижской Академіи Наукѣ, Апрѣля 24 числа 1754 года.

Ужасная и свирѣпая болѣзнь, коей сѣмена носимѣ мы вѣ крови, истребляетъ, извлеченіемъ, и безобразитъ четвертую часть человѣческаго рода. Кара будучи старагосвѣта, опустошила она новой болѣе, нежели мечь завоевателей онаго. Она оружіе смерти поражающее, не разбирая лѣтъ, полу, чина и обиталища. Мало семей избѣгаютъ горестной дани требуемой ею. Наипаче она вѣ городахъ и при дворахъ самыхъ великолѣпныхъ чинитъ опустошенія. (a) Чѣмъ выше и драгоцѣнѣе главы, коимъ она

(a) Особливое либо состояніе воздуха, либо разность кушанія, или что другое причиною, что бываетъ обыкновенно оспа гораздо опаснѣе по городамъ, а особливо у взрослыхъ людей и дѣтей нѣжно воспитанныхъ.

оз. въ
Т.
г͡. АблабовЛА ОЛНА
вфво
о прлолоанїи воєпы-
сттанаа. имъ. бамимъ. при полни.
номъ. оованїи. королєєско а.
Елжєкои. авадемї нокъ, ап͡ръ.
лаоа. тгла. ечва года.
Ожаєнаа в͡євиеѢпаа. болъзнь, коєв. ф исна
и осимъ ны ъсрови; и стрьолаєтъ. изф.
vвфвиваєтъ, и обоооразитъ. пстьє отоw пасть
пбло бъ͡вє гваго. роза. Кава. оаоупи. ставаго
вф та, о побтотила она. но сок.оолъє, и єже.
ли. иечь. завоєватьлев. онаго. Вна. ооужїє
биєрти. поражавшее, и є͡ разоираа. лѢтъ, полоу, чина. и овталиша. Мало. є͡є мвв и в.

Upload another image
Copy to clipboard

Анализировать насыщенные ошибками тексты так, как если бы это были исправные тексты, некорректно. Некоторые специалисты высказывали идеи, которые бы позволяли работать с плохо распознанными текстами[1], но пока что это направление остается в области теоретических рассуждений.

Ошибки нужно исправлять, и делать это приходится вручную. Здесь нет хороших программных решений, хотя перспективным представляется спелл-чекер для текстов в старой орфографии,

[1] Jiang M. et al. Impact of OCR quality on BERT embeddings in the domain classification of book excerpts // Proceedings http://ceur-ws. org ISSN. 2021. Т. 1613. С. 0073.

разработанный в рамках учебного проекта в школе лингвистики НИУ ВШЭ[1]. Вычитка распознанных текстов — самая затратная (человеческий труд обычно стоит больших денег, чем машинное время) и нетворческая часть исследовательской работы.

В текстах из интернета мы можем столкнуться с тем, что вместо привычных нам символов обнаружим т.н. HTML-entities (например, « вместо «). В чем опасность того, что исследователь не позаботится об удалении HTML-entities? Релевантное слово склеится со словом «laquo», и мы неправильно посчитаем частотность первого. Это может повлиять на наши выводы. Допустим, сразу после слова «map» (‘карта’) в тексте будет следовать закрывающая кавычка: *map«*. Наш код предобработки удалит знаки амперсанда и точки с запятой, получится квазислово *maplaquo*. Таким образом, подсчитывая частотность слова «map», мы упустим его вхождение в связке с HTML-entity ««».

В процессе предобработки необходимо превратить HTML-entities в соответствующие им символы. В Python это можно сделать следующим способом:

```
import html
html_decoded = html.unescape(html_string)
```

Кроме того, на странице как минимум два раза представлено название текста, имя автора: первый раз они отражены в части, которая называется header, например, в теге <title>, содержимое которого выводится в заголовке окна браузера; второй раз в части страницы, которая называется <body> и видима для пользователя в рабочей области окна. Это тоже может повлиять на последующее вычисление частотности слов. Если мы не удалим навигационные элементы сайта, то самым частотным словом у нас может стать слово «главный»: на веб-страницах обычно есть ссылка, ведущая на *главную* страницу сайта.

Для исследователя полезно, чтобы текст, с которым он работал, был структурированным. Как и в случае с обвязкой сайта, важно разделять в тексте основное его тело и заголовок. Это позволит в дальнейшем не путаться в распределениях лексики и не допускать перекоса при составлении частотного словаря. Например, известно, что название романа Стендаля «Красное и черное» никак

[1] https://github.com/dhhse/Otechestvennie_zapiski/tree/master/kate_data

не подкрепляется в самом тексте упоминанием этих цветов[1]. Понятия, стоящие за этими цветами, важны, но сами цвета при этом выражены именно в заглавии. Существуют и исследования, опирающиеся при этом только на данные анализа заглавий без обращения к тексту романов[2]. Если текст большого романа будет разбит на главы, то исследователь получит возможность выяснить, чем эти главы с точки зрения статистики данных отличаются друг от друга.

Такие исследования могут проводиться не только на романах. В статье «Почему ошибался Жуковский»[3] речь идет о метрических «сбоях» в переводе «Одиссеи», то есть о таких строках, в которых поэт использовал семь или пять стоп вместо требуемых шести. Такие строки встречаются только в нескольких песнях поэмы, и эти песни противопоставлены «безошибочным». Благодаря структуризованности текста, выделенности границ песен, оказалось возможным провести все нужные подсчеты и сделать вывод о том, что метрические аномалии появляются главным образом в «морских» песнях поэмы и отсутствуют в «сухопутных».

Но для предобработки текста существенны и более мелкие членения. Так, для некоторых типов аналитических операций важно бывает разбить текст на отдельные предложения. Предложение — это минимальный смысловой контекст, опираясь на который мы можем смоделировать значение слова.

Не всегда ясно, как простым способом произвести это членение. Вроде бы у нас есть в тексте для этого хороший маркер — знак конца предложения. Это точка, а в некоторых случаях — восклицательный или вопросительный знак. Но есть и множество случаев, когда все не так просто. Например, та же точка используется и для сокращения, после инициалов. Слово «Вячеслав» сокращается до «Вяч.» и если бы машина решила, что точка всегда разделяет предложения, то граница предложений прошла бы здесь между именем и отчеством человека: «Заглядывает “в башню” Вяч. Иванова, когда там водят “хороводы”

[1] Лотман Ю. М. Несколько слов к проблеме «Стендаль и Стерн» (Почему Стендаль назвал свой роман «Красное и черное»?) // Лотман Ю. М. Избранные статьи: в 3 т. Т. 3. Таллинн, 1993. С. 428–429.

[2] Моретти Ф. Корпорация стиля: размышления о 7 тысячах заглавий (британские романы 1740–1850) // Моретти Ф. Дальнее чтение / пер. с англ. А. Вдовина, О. Собчука, А. Шели; науч. ред. перевода И. Кушнарева. М.: Изд-во Института Гайдара, 2016. С. 248–287.

[3] Орехов Б. В. Почему ошибался Жуковский: о внутритекстовых причинах метрических сбоев в «Одиссее» // М. Л. Гаспарову — стиховеду. In Memoriam / сост. М. В. Акимова, М. Г. Тарлинская. М.: Языки славянской культуры, 2017. С. 73–89.

и поют вакхические песни, в хламидах и венках» [З. Н. Гиппиус. Задумчивый странник (о Розанове) (1923)].

Инициалы — это не единственный случай такого рода трудностей. Вообще-то в русской типографике не принято ставить точку после сокращения «миллион», но в реальных текстах (особенно из интернета) мы, разумеется, встретимся с тем, что точка в этом месте ставится. Поэтому разделение на предложения должно быть устроено более умным способом, и для этого есть готовые программные решения. О них см. ниже.

Разбивать тексты нужно и на более мелкие единицы, например, «токены». Токен — это и слово, и знак препинания. Например, слово «так» в некоторых контекстах существует вместе со следующей за ним пунктуацией:

Так!.. Но, прощаясь с римской славой,
С Капитолийской высоты
Во всем величье видел ты
Закат звезды ее кровавый!..

Важно разделить их и представить отдельно, потому что в противном случае у нас будет две единицы для подсчета — слово «так» с «прилипшей» к нему пунктуацией и слово «так» без пунктуации. Нас почти наверняка будет интересовать объединенная частотность слова «так» в обоих этих случаях, поэтому пунктуацию нужно отрезать.

В Python есть переменная, содержащая почти все нужные нам небуквенные символы, которые имеет смысл отрезать от слов.

```
from string import punctuation, digits
punct = punctuation + '""--…“”\n\t’ + digits
```

Если мы работаем не с русским языком, а с какими-нибудь нестандартными для европейского культурного пространства письменностями, то ситуация еще сложнее. Например, китайская или тайская графика вообще не подразумевают деления на слова с помощью пробелов. Так выглядит первая строфа перевода на китайский язык стихотворения Ф. И. Тютчева «Silentium!»:

别声张，要好好地收起
自己的感情，自己的向往；

任凭它们在心灵深处
默默地升起，悄悄地沉落，
像繁星，在夜空中
任你观赏，可别声张！

Носитель языка при чтении легко находит границы слов, а для компьютера это проблема.

Наконец, если в языке слова изменяются (например, по падежам — как в русском языке), значит, перед нами будет стоять задача лемматизации, то есть автоматического нахождения леммы, проще говоря — словарной формы слова. Скорее всего, частотный словарь, который мы хотим получить, это словарь именно лемм, а не конкретных словоформ, из которых состоит текст. Иначе говоря, нас будет интересовать не частотность отдельных форм «окнами, окна, окну», а частота всех их сразу.

Хороший инструмент для той самой умной сегментации текста на предложения (или, как еще говорят, сплиттинга, от слова *split* — разрезать) — это программная библиотека natasha[1], написанная на языке программирования Python. Сплиттингом ее функциональность не ограничивается, но нас сейчас интересует именно он.

```
from natasha import (
Segmenter,
Doc
)

segmenter = Segmenter()
doc = Doc(text)
doc.segment(segmenter)
```

Библиотека умеет разбивать текст на предложения и предложения на отдельные токены. Числовые значения у параметров «старт» и «стоп» — это номер символа, на котором начинается или заканчивается соответствующее предложение в тексте.

Структурирование текста позволяет получать информацию автоматически. Структуру фиксирует разметка, то есть система специальных указаний для компьютера на то, чем являются те или иные сегменты текста. Такая разметка может существовать, например,

[1] https://github.com/natasha/natasha

в виде XML-тегов. Так структурирован текст стихотворения в поэтическом корпусе НКРЯ:

```
<?xml version="1.0" encoding="utf-8"?>
<html><head>
<title>Тютчев Ф.И. Какое дикое ущелье!.. (1835)</
title>
</head>
<body>
<p class="verse"><line meter="Я4ж"/>Какоѐ дѝкоѐ
<rhyme-zone/>ущѐлье!<br/>
<line meter="Я4м"/>Ко мнѐ навстрѐчу клю̀ч <rhyme-
zone/>бежи́т —<br/>
<line meter="Я4ж"/>Он в доль̀ спеши́т на <rhyme-
zone/>ново̀сѐлье, —<br/>
<line meter="Я4м"/>Я лѐзу ввѐрх, где ѐль <rhyme-
zone/>стои́т.</p>
<p class="verse"><line meter="Я4ж"/>Вот взобра̀лся̆
я на̀ <rhyme-zone/>верши́ну,<br/>
<line meter="Я4м"/>Сижу̀ здесь ра̀достѐн и <rhyme-
zone/>ти́х —<br/>
<line meter="Я4ж"/>Ты к лю̀дям, клю̀ч, спеши́шь в <rhyme-
zone/>доли́ну, —<br/>
<line meter="Я4м"/>Попробу́й, ка̀ково̀ у <rhyme-
zone/>ни́х!</p>
<p class="date"><noindex>&lt;1835&gt;</noindex></p>
</body></html>
```

Здесь заглавие отделено от собственно текста. Это позволяет отдельно подсчитать частотность слов, входящих в поэтическую строку и оставшихся за ее пределами, связать конкретные лексемы и метр строки[1]. С помощью тега <p> отделены друг от друга строфы. Выделены и рифмующиеся слова: это тег <rhyme-zone>.

Если для наших аналитических целей мы составляем частотный словарь, то больше информации нам даст именно подсчет лемм, а не отдельных словоформ. Распределение по словоформам часто

[1] Подробнее см.: Orekhov B. Lexis Meets Meter: Attraction of Lexical Units in Russian Verse // CLLS2016. Computational Linguistics and Language Science. Proceedings of the Workshop on Computational Linguistics and Language Science. 2017. Vol-1886. P. 110–121.

имеет случайный характер, и информация о частотах словоформ «окнами» и «окнах» нам скорее всего ничего полезного не даст, а вот если мы будем знать частотность всех этих форм, то эта информация уже будет иметь системный характер для, например, тематики текста.

Если мы работаем с большим текстом или даже большой коллекцией текстов, то вручную для каждого «окнами» словарную форму не пропишем — это будет слишком большая и бессмысленная работа. Существуют программные решения для лемматизации текстов. В простых случаях они работают эффективно, обычно это частотные и регулярно образующие свои формы слова вроде «земля» (земли, земле, землей и под.), «дерево» (дерева, дереву, деревом) и подобных. Но есть и сложные случаи. Во-первых, это имена собственные, особенно пришедшие из других языков. Их облик затрудняет машине распознавание способа формоизменения. Во-вторых, это формы, которые теоретически могут восходить к разным леммам, и без контекста будет непонятно, к какой именно. Например, слово «мыла» может быть и глаголом «мыть» в форме прошедшего времени, и существительным «мыло» в родительном падеже. Только в контексте «мама мыла раму» мы поймем, что речь именно о глаголе. Такие трудности называются грамматической омонимией. Некоторые программы умеют «смотреть» на контекст и реконструировать наиболее вероятную для такого контекста лемму. Такое умение называется «снятием омонимии».

Одно из лучших решений для автоматической лемматизации — программа под названием mystem[1]. Она умеет и восстанавливать словарную форму слова, и снимать омонимию. Программа доступна в виде бинарного исполняемого файла для разных операционных систем, но существует и обертка на языке Python[2].

Другой популярный инструмент для лемматизации — это библиотека pymorphy2[3], она написана на чистом языке Python, но не умеет учитывать контекст и снимать омонимию. В омонимичных случаях программа выдает весь возможный набор вариантов в порядке от наиболее частотного в языке к наименее частотному. Естественно, что исследователю может не повезти и тогда программа выдаст частотный, но неподходящий для данного контекста вариант.

---

[1] Зобнин А.И., Носырев Г.В. Морфологический анализатор MyStem 3.0 // Труды Института русского языка им. В.В. Виноградова. 2015. Т. 6. С. 300–310.

[2] https://pypi.org/project/pymystem3/

[3] https://pymorphy2.readthedocs.io/

Эти программы умеют не только восстанавливать словарную форму, но и делать морфологический разбор, то есть сообщать пользователю, к какой части речи принадлежит слово, в какой грамматической форме оно стоит (то есть в каком падеже или форме какого времени и лица).

Число таких инструментов множится. Лемматизировать текст можно и с помощью той же natasha, которая упоминалась выше.

```
for token in doc.tokens:
token.lemmatize(morph_vocab)
for _ in doc.tokens:
print(_.lemma)
```

Как мы уже говорили, программы неизбежно ошибаются. Имена собственные, которые похожи на какие-то косвенные формы русских слов, могут лемматизироваться, исходя из ложных оснований. Например, фамилия бывшего президента Франции Николя Саркози напомнила программе форму русского повелительного наклонения вроде «своди», «замени». От этих форм восстанавливается лемма «сводить», «заменить», а от «Саркози» — «Саркозить». Это немного напоминает языковую игру, в которой слово «крокодил» рассматривается как глагол: «крокодил, крокожу и буду крокодить». Американский политический деятель Дик Чейни рассматривается программой как родительный падеж от существительного «чейня», а сокращение «Изд.» (то есть «издательство») как родительный падеж множественного числа от слова «изда».

Частотность слова

Частотность слова — это ценный ресурс, с которым полезно работать и филологу, и тому, кто осуществляет цифровое исследование художественного текста. Полезность этого параметра доказывается и его использованием в разнообразных технологиях обработки текста, о которых мы скажем несколько позже. Частота слова — это тот самый «крючок», за который может зацепиться компьютер в подходах к «пониманию» информации, которая содержится в тексте.

Частотные словари — это по сути один из способов моделирования текста. Моделирование — это научное упрощение объекта,

попытка представить его в более общем виде, чем в реальном бытовании. Через частоту слов можно реконструировать и тематику текстов, и их стилистические особенности. Но при этом не следует трактовать частотный словарь слишком прямолинейно, то есть предлагать простые и наивные интерпретации частотных списков. Существует множество разнообразных факторов, которые влияют на частоту появления слова в тексте, но при этом плохо отражаются в словаре.

Существуют старые, выходившие еще в 1970-е, исследования частотности слов в поэзии, например, за авторством Гейра Хьетсо. Этот скандинавский филолог-русист стал особенно известен благодаря инициированному им количественному исследованию романа «Тихий Дон», позволившему, как казалось автору, утверждать, что авторство этого текста принадлежит Шолохову[1]. В том исследовании можно обнаружить заметное количество методологических ошибок, не позволяющих считать тему исчерпанной. Более подробно этот случай разбирается в более современной публикации, посвященной тому же предмету[2].

По данным Хьетсо[3], у Баратынского первые места в частотном словаре занимают слова «мятежный» и «счастливый». Как кажется интуитивно, ни то ни другое слово не отражают особенности поэтики Баратынского. У Тютчева на первое место попадает слово «великий», что тоже довольно далеко от главных для нас мотивов в творчестве этого поэта[4]. Еще раз подчеркнем, что к интерпретации частотности нужно подходить осторожно. Это важный способ описания текста, но далеко не универсальный.

Мы не всегда знаем, частотность отражает распространенность слова во всех текстах нашей коллекции или только внутри одного текста. Частотное слово может получить такой статус благодаря просто повторению, скажем, в рефрене одного поэтического текста. Тогда, конечно, это случайный «выброс», и служить характеристикой

---

[1] Хьетсо Г., Густавссон С., Бекман Б., Гил С. Кто написал «Тихий Дон»? (Проблема авторства «Тихого Дона») / пер. А. В. Ващенко, Н. С. Ноздриной. М.: Книга, 1989.

[2] Великанова Н. П., Орехов Б. В. Цифровая текстология: атрибуция текста на примере романа М. А. Шолохова «Тихий Дон» // Мир Шолохова. Научно-просветительский общенациональный журнал. 2019. № 1. С. 70–82.

[3] Хетсо Г. Лексика стихотворений Лермонтова. Опыт количественного описания. Oslo, 1973. 44 с.

[4] Орехов Б. В. Принципы организации мотивной структуры в лирике Ф. И. Тютчева: автореф. дис. … канд. филол. наук. Воронеж, 2008.

всего набора текстов этот выброс не может. Поэтому было бы неправильно на основе таких подсчетов говорить, что частотность репрезентирует художественный мир того автора, количественное исследование текстов которого мы проводим.

На распределение частотностей слов влияет и жанр. Если, к примеру, мы обозреваем жанр писем, то частотными у нас окажутся слова, которые входят в традиционные формулы приветствия и прощания, повторяемые в письмах: «Привет», «Будь здоров», «Всего доброго». Это тоже не черта авторского стиля, а особенность жанра, заставляющая автора писать так, а не иначе. Аналогичным образом влияет на частотность и тематика текста, которая в большей степени будет отражать не авторскую стилистику, а законы построения художественного мира. На частотности отразиться то, на космическом корабле происходит действие романа или в провинциальном селе.

Важно помнить, что самыми частотными в тексте всегда будут служебные слова. Первые 10–20, а то и 30 слов в частотном словаре, ранжированном от большего к меньшему, это всегда предлоги, союзы, частицы[1]:

1. и
2. в
3. не
4. на
5. я
6. быть
7. он
8. с
9. что
10. а
11. по
12. это
13. она
14. этот
15. к

Причем это не особенность исключительно русского языка и не особенность какого-то конкретного автора, а универсальная закономерность.

[1] Ляшевская О.Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

Чтобы данные в разных частотных словарях были сопоставимы, то есть чтобы их можно было сравнивать, частотность нужно измерять не только в абсолютных числах, но и в единицах IPM (instance per million). Абсолютное значение — это, например, 5, или 15, или 45 — сколько раз некоторое слово встретилось в нашем корпусе. IPM — это та же самая частотность в расчете на один миллион словоупотреблений. Это значит, что если длина нашего корпуса равна миллиону слов, то IPM для слова, встретившегося в таком корпусе один раз, будет 1, три раза — 3 и т.д. А если длина нашего корпуса два миллиона, то IPM для абсолютной встречаемости 1 будет 0,5, для абсолютной частотности 3 будет 1,5. Это позволит нам получить такие цифры, которые дадут возможность для сравнения — того же Тютчева с Баратынским или Лермонтовым, даже если объем их текстов будет меньше миллиона слов.

Важно помнить, что частотный словарь всегда подчиняется так называемому закону Ципфа. Закон Ципфа — это частный случай того, что в математике называется степенным распределением. Если грубо упростить, то степенное распределение означает, что частотность первых в списке слов будет очень высокой, но станет быстро падать. Примерно четверть всех слов из словаря будет употреблена только два раза, а половина всех слов будет иметь частотность 1. Важно помнить, что это не особенность какого-то одного текста или автора. Так будет всегда для текстов, созданных на естественном языке. Если озвученное выше правило с наиболее частотными служебными словами, или закон Ципфа, не выполняется, это значит, что с текстом или нашими подсчетами «что-то не так». Например, в некоторых языковых разделах Википедии есть не написанные человеком, а сгенерированные по определенному шаблону статьи. Чтобы их обнаружить, достаточно составить частотный словарь этой Википедии и увидеть там в десятке наиболее частотных слова вроде «река», «бассейн» и подобные[1]. Это будет значить, что Википедия состоит в том числе и из десятков тысяч единообразных статей о реках и других водных объектах. Это уже не вполне естественный язык и не вполне естественный набор текстов.

Один из способов посчитать частотность правильно — это соотнести ее с частотностью того же самого слова в других документах

[1] Орехов Б.В., Решетников К.Ю. Государственные языки России в Википедии: к вопросу о сетевой активности миноритарных языковых сообществ // Настройка языка: управление коммуникациями на постсоветском пространстве: коллективная монография. М.: Новое литературное обозрение, 2016. С. 263–281.

текстовой коллекции. Это важно для тех случаев, когда нам важно понять, что частотность информативна и ее можно привлекать к научному анализу. Речь идет о метрике, которая называется TF-IDF. Компьютерная лингвистика — это дисциплина, которая занимается подсчетами в текстах и строит на основе выявленных закономерностей технологии обработки больших массивов текстовой информации.

Одна из задач, которые решают компьютерные лингвисты, — это выделение в тексте ключевых слов. Эти слова способны представлять весь смысл документа в сжатом виде, они выделяют для нас в тексте главное. Их поиск можно алгоритмизировать, если принять, что они часто встречаются в некотором одном интересующем нас документе, но при этом редко во всех остальных. Такой подход кажется очень осмысленным.

Что стоит за этой метрикой? Если мы возьмем абстрактный текст про Францию, ключевыми словами будут сама Франция и Париж, потому что они встретятся в этом тексте, а в текстах про другие страны или вообще не про страны Франция и Париж встречаются реже. Но все-таки встретятся, потому что это известная страна и известная ее столица.

В статье о Франции из Википедии, разумеется, есть и слово «Франция», и слово «Париж».

Но в другой статье из Википедии, не про Францию, а про русского поэта Владислава Ходасевича, видно, что несмотря на то, что текст в целом посвящен другому предмету, слова «Франция» и «Париж» тоже присутствуют. Что нам нужно сделать, чтобы понять, является ли для данного текста слово «Франция» ключевым? Правда ли слово «Париж» выражает важную информацию из текста, с которым мы имеем дело?

Для вычисления того, насколько интересующие нас слова являются ключевыми, и существует метрика tf-idf. Если сказать упрощенно, то tf — это сокращение от term frequency, то есть это значение абсолютной частотности, с которой некоторое слово встретилось в тексте или документе. Текст и документ — это в терминах компьютерной лингвистики одно и то же.

Idf — вторая часть этого термина — это inverted document frequency. То есть насколько редко это слово встречается во всех остальных документах. С помощью этого параметра мы можем вычислить не просто частотность слова, но и понять, насколько эта частотность важна на фоне остальных текстов.

Нам совершенно не обязательно пытаться самостоятельно записать формулы вычисления метрики в коде, потому что существуют уже готовые библиотеки на Python, которые включают в себя функционал вычисления нужных нам значений. Приведем код для вычисления idf с помощью библиотеки sklearn.

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
texts[0]
```

```
'купаться в шторм запрещать . \n заплывать -- не возвращаться . \n волна накатный бревно \n расплющива
ть бедный артист ! \n но среди бешеный вал \n быть тихий волна -- / пасат ,\n как среди гром к
аблук \n стопа / неслышный / босой . \n ты от берег влечь \n не удалой бесшабашность ,\n а ужасать
расчет --\n в открытый море / безопасный . \n артист , над мировой волна \n ты носиться от жизнь
к смерть ,\n как ограниченный дуга \n латунный / сгорбленный / рейсфедер ! \n но слышать зоркий спин
а \n среди безвыходный сальто ,\n как зарождаться волна \n с протяжный имя -- пасат . \n « пасат , во
звращать волна , пасат ,\n запретный мой заплыв , но хлынуть тишина возврат ,\n я обожать вода --
/ но что она без земля ? / пустой ! \n я обожать свобода -- / но что она без любовь , / пас
ат ! \n нести я , пока носить , оставлять на берег , -- быть святой ! \n я вставать / и , пошатыва
ться , / ты поблагодарить ,\n но ты растворяться в море , / не поглядеть , / пасат … »\n'
```

```
пасат 0.5448942548714197
волна 0.3863655788838256
среди 0.18024593980104597
артист 0.16964603222642136
но 0.16143537093042773
обожать 0.13717149335466636
берег 0.13405707874582715
море 0.11356792593812406
ты 0.09958267454732944
поблагодарить 0.09081570914523662
пошатываться 0.09081570914523662
возврат 0.09081570914523662
хлынуть 0.09081570914523662
```

```
tfidf_vectorizer = TfidfVectorizer(
max_df=0.95, min_df=2, max_features=n_features
)
```

Здесь мы видим результат вычислений с помощью кода sklearn. Это одно из стихотворений Андрея Вознесенского[1]:

Купаться в шторм запрещено.
Заплывшему — не возвратиться.
Волны накатное бревно
расплющит бедного артиста!

[1] См. также: Орехов Б. В. Метрическое и лексическое разнообразие в стихах А. А. Вознесенского // Труды Института русского языка им. В. В. Виноградова. 2022. № 3. С. 50–58.

Слова уже лемматизированы, а значение idf ранжировано для корпуса его ранних произведений. Мы видим, что такие слова, как «пасат» и «волна» значимы именно для этого стихотворения, а в остальных текстах раннего периода эти слова или встречаются один раз, или не встречаются вовсе, что показывает их ключевой характер для текста, с которым мы работаем.

Совместная встречаемость слов

Мы разобрали случаи, когда частотный словарь способен подсказать аналитику что-то о том, как устроен текст, на котором этот частотный словарь был построен. Единицами частотного словаря по умолчанию являются слова. В то же время многие частотные слова попадают в этот класс случайно, и в целом интерпретация частотного списка требует аккуратности. Чтобы избежать традиционных ловушек, связанных с частотностью, компьютерная лингвистика разработала несколько более умных подходов, нивелирующих влияние случайных факторов на позицию в частотном списке.

Но еще более важным количественным источником информации о тексте является совместная встречаемость слов. К сожалению, и здесь мы не можем быть уверены, что данные о совместной встречаемости, с которой мы работаем, не случайны. В одном тексте могут оказаться какие угодно слова, и их сочетание не будет иметь никакого семантического наполнения. Поэтому и для количественной оценки неслучайности совместного появления лексических единиц в тексте есть компьютерно-лингвистические методики.

Слова, появляющиеся вместе в тексте не случайно, могут отражать тему, которой посвящен текст. Мы уже видели это на примере статей Википедии о Франции. Слова, появляющиеся вместе в тексте, могут отражать особенности авторского стиля, то есть сознательное или бессознательное стремление к такому словоупотреблению. Обычно стилистика особенно ярко отражается на так называемых n-gram'ах. Gram — это единица текста. В зависимости от задач исследователей «грамом» может быть и буква, и слово. Мы далее будем говорить именно о словах, хотя совершенно естественны и такие контексты, в которых под «грамом» говорящий будет понимать и более мелкую единицу письма. N — это буква,

означающая неопределенность, переменную. На место N можно подставить какое-то число, например двойку, и тогда речь будет идти о биграмме — комплексе из двух стоящих друг за другом слов. Но можно пойти дальше и работать с триграммой, тетраграммой, то есть последовательностью, состоящей из трех или четырех слов.

N-gram'ы важны для разных исследований текстов, они являют собой частный случай совместной встречаемости слов, такую разновидность, когда имеют в виду именно стоящие друг за другом слова, которые не разрываются дистантным расположением в тексте.

N-gram'ы — это не случайные совпадения; если одни и те же слова часто стоят рядом друг с другом, это отражает в том числе и авторскую стилистику.

Еще один важный в контексте этого разговора термин — коллокации, то есть такие n-gram'ы, которые встречаются вместе в тексте значимо часто по сравнению с их индивидуальными частотами. Кроме того, иногда говорят о том, что коллокации обязательно должны представлять собой грамматически связанные слова, а не просто случайные. То есть «золотой осени» — это коллокация, а «вода серебряный» уже нет.

Что значит, что слова встречаются вместе в тексте статистически значимо по сравнению с их индивидуальными частотами? Это значит, что мы сможем судить о том, случайно или не случайно слова оказались рядом, только сделав некоторые вычисления, в которых будут учтены частотности слов, входящих в возможную коллокацию, по отдельности. Одна из метрик, которая дает нам представление о том, с чем мы имеем дело, это PMI, то есть piecewise mutual information, «покомпконентная взаимная информация». Не стоит путать ее с PMI, то есть instance per million, единицей, в которых следует отражать частотность частотного словаря. О ней мы говорили раньше.

PMI показывает «необычность», «непредсказуемость» явления, состоящего в том, что происходит сразу несколько событий одновременно, и у каждого из этих событий по отдельности есть своя вероятность. Хитрость в том, чтобы оценить их совместную вероятность. PMI показывает, насколько сильна связь в слов в сочетании, вне зависимости от частности самих слов. То есть один и тот же показатель PMI могут иметь две комбинации, при этом одна из них составлена из низкочастных слов, а другая из высочастных. Тогда PMI показывает, насколько часто два слова объединяются вместе относительно их собственных частот. Коллокация с высоким

PMI — это редкое, неожиданное явление, коллокация с низким PMI — обычное, частое, предсказуемое, рядовое явление. Важно при этом принять во внимание частотности самих слов в анализируемом корпусе.

Как и в других случаях, нам не обязательно воспроизводить эту формулу самостоятельно, поскольку существуют готовые программные решения. PMI можно вычислить с помощью соответствующей функции внутри программного пакета для Python под названием NLTK, то есть natural language toolkit, «набор инструментов для естественного языка».

Вот результат вычисления PMI для произвольного текста:

4,486 Федор Достоевский
4,481 Иван Тургенев
4,481 многолетний растение
—
2,376 велосипед ветка
1,4875 задание колодец

Высокое значение PMI оказалось у слов, которые действительно часто встречаются вместе, поскольку являются именем и фамилией известных писателей. Еще один пример высокого PMI — это коллокация «многолетнее растение». Поскольку в тексте уже проведена лемматизация и словарной формой для прилагательного в русском языке является форма мужского рода, то в списке словосочетание выглядит несогласованным по роду. Но это просто результат препроцессинга текста.

Под чертой — ряд сочетаний слов с низким PMI. Это слова, совместное появление которых в тексте выглядит случайным и не сообщает полезной информации. Важно понимать, что значимость совместной встречаемости слов в тексте можно оценивать не на глазок, а с помощью конкретных вычислительных методик.

В области цифрового литературоведения совместная встречаемость слов (называемая лексическими комбинациями) стала основой для ряда исследований смоленской филологической школы[1].

[1] См., например: Павлова Л. В., Романова И. В. Лексические комбинации в «Кормчих звездах» Вячеслава Иванова (из опыта применения компьютерного комплекса «Гипертекстовый поиск слов-спутников в авторских текстах») // Новый филологический вестник. 2019. № 3 (50).

Тематическое моделирование

Одна из технологий, которые базируются на идее совместной встречаемости слов как информационного ресурса, называется тематическим моделированием (topic modeling). Тематическое моделирование применяется в информационном поиске, то есть встроено в поисковые машины, в анализе новостного потока и других практических приложениях компьютерной обработки текста.

Но тематическое моделирование хорошо вписывается и в концептуальную рамку digital humanities. Особенно удачно подходит для осмысления места и вклада этой технологии в цифровые гуманитарные науки понятие distant reading, которое на русский язык переведено как «дальнее чтение»[1]. Термин придуман американским литературоведом итальянского происхождения Франко Моретти. Моретти назвал так свою книгу, под обложкой которой собрал несколько важных для последующего развития digital humanities эссе. «Дальнее чтение» (или, как его еще можно передать по-русски: «отвлеченное чтение») — это понятие, не имеющее точного и строгого определения, то есть так называемый «зонтичный» термин, удобный для обозначения с его помощью разных подходов, которые все же имеют между собой что-то родственное. В данном случае то, что объединяет разные подходы, это оппозиция традиционному медленному чтению (по-английски — close reading). При медленном чтении в классическом литературоведении внимательное изучение текста совершается самим исследователем, который пытается интерпретировать объект своего интереса. Distant reading — это попытка изучать текст, не читая его. Как это возможно? Между текстом и исследователем появляется новая сущность, научная модель, создаваемая компьютером. Машина по определенной программе извлекает из текста важные для его изучения параметры и представляет их исследователю в сжатом виде. Этот вид можно, например, визуализировать, то есть отрисовать на графике, или представить в виде таблицы. Такими параметрами могут быть и частотные списки слов, и система взаимодействий между персонажами, и даже длина заглавия произведения. Именно с этой промежуточной моделью исследователь и работает, не имея необходимости заглядывать в исходный текст.

[1] Моретти Ф. Дальнее чтение / пер. с англ. А. Вдовина, О. Собчука, А. Шели; науч. ред. перевода И. Кушнарева. М.: Изд-во Института Гайдара, 2016.

Лучший способ понять тематическое моделирование в контексте digital humanities — представить его в виде одного из подходов distant reading, поскольку тематическое моделирование не подразумевает чтения человеком текстов, к которым моделирование применяется.

Как можно догадаться благодаря названию, тематическое моделирование — это технология, которая позволяет выделять темы текста, то есть, использовав компьютер, узнать, о чем текст или текстовая коллекция. Правда, это не значит, что темы предстают в привычном нам со школы виде, где учителя предлагают нам написать сочинения на тему природы, родины или любви. Темы в рамках topic modeling — это ряды слов, которые заранее никак не озаглавлены. Машина, которая реализует технологию тематического моделирования, может представить нам список слов по типу «финансы, деньги, транш, вклад, счет, банк». И уже человек, а не сама машина может увидеть в этом списке объединяющее начало и присвоить ему ярлык, например «тема финансов».

Как получаются эти «темы» или, вернее сказать, списки слов? Фактически слова, образующие темы, — это слова, которые встречаются в текстах вместе. Так реализует себя в конкретной технологии концепция совместной встречаемости слов как информационного ресурса. Но в данном случае речь не идет о простых биграммах или даже коллокациях. Слова, объединяемые в темы, могут располагаться в тексте дистантно, и, чтобы определить их взаимное тематическое тяготение, используется специальный математический аппарат.

У нас есть два наиболее распространенных алгоритма со своими преимуществами и недостатками, которые обычно используются для тематического моделирования. Самый популярный алгоритм, который хорошо подходит для прозаических повествовательных текстов, это LDA (Latent Dirichlet allocation, Латентное размещение Дирихле). Второй используется реже, но хорошо себя показывает на поэтических текстах: NMF (Non-negative Matrix factorization, неотрицательное матричное разложение). Оба алгоритма подразумевают составление так называемой терм-документной матрицы, то есть, упрощенно говоря, таблицы, в которой прописано, в каком тексте какое слово и с какой частотой встретилось. Затем эта таблица подвергается математическим преобразованиям. Но, как и в прочих случаях, нам не нужно пытаться реализовать эти алгоритмы самостоятельно, так как они уже встроены в разнообразные программные библиотеки Python: sklearn и gensim.

В работе со стихами специалисты обычно говорят, что имеет смысл предпринимать максимально дробное членение текстов, желательно по одной строфе, а не анализировать стихотворение целиком[1].

У тематического моделирования уже есть несколько имеющихся применений в digital humanities. Одна из известных работ на эту тему — это анализ 45 тысяч стенограмм французского парламента, который начал работать в конце XVIII века[2]. Исследователей интересовало, какие темы в выступлениях депутатов парламента обладают, скажем так, «наибольшей живучестью». Иными словами, какие темы продолжают обсуждаться, а какие, наоборот, затихают и не получают продолжения в дискуссиях. Оказалось, что наиболее интересными для последующих обсуждений оказывались темы, которые первоначально поднимались депутатами от левых фракций. Темы, которые предлагались правыми депутатами, были не очень долговечными.

Еще один пример применения тематического моделирования — это анализ классической французской драмы в статье, опубликованной в журнале *Digital Humanities Quarterly*[3]. В этой работе показано, что тематика комедий и трагедий во французской драматургии очень сильно отличается. Если попробовать отрисовать анализировавшиеся тексты на графике, что и сделано в статье, то видно, что комедии, которые обозначены красными точками, находятся довольно далеко от трагедий, обозначенных синими точками. Видно также, что есть и немногочисленные тексты, которые близки по тематике и кластеру комедий, и кластеру трагедий, но в целом это все же сильно различающиеся классы. Нечто среднее представляют собой трагикомедии. Это как минимум свидетельствует о том, что тематическое моделирование хорошо отвечает заявленному предмету, то есть в самом деле позволяет увидеть тематическую разницу между жанрами.

---

[1] Haider T., & Eger S. (2019). Semantic change and emerging tropes in a large corpus of new high German poetry. In N. Tahmasebi, L. Borin, A. Jatowt, & Y. Xu (Eds.). Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change (pp. 216–222). Association for Computational Linguistics, 2019

[2] Alexander T. J. Barron, Jenny Huang, Rebecca L. Spang, and Simon DeDeo Individuals, institutions, and innovation in the debates of the French Revolution // PNAS May 1, 2018 115 (18) 4607–4612.

[3] Schöch C. Topic modeling genre: An exploration of French classical and enlightenment drama // Digital Humanities Quarterly. Vol. 11. No. 2. 2017. URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Можно проводить исследования с помощью тематического моделирования и на русскоязычном материале. При этом важно учитывать то, о чем мы говорили раньше: нужна правильная обработка текстов, лемматизация, избавление от знаков препинания. В данном случае лемматизация особенно важна, так как носителем темы текста является не конкретная словоформа в косвенном падеже, а именно лемма.

Пример применения тематического моделирования на русскоязычном материале — статья Лейбова и Орехова о поэтической топике Крыма[1]. Здесь исследуются многочисленные, доходящие до десятков тысяч, стихотворения, посвященные Крыму, написанные непрофессиональными поэтами, публикующимися на сайте stihi.ru. Прочесть все это текстовое богатство глазами было бы невозможно, а исследовательская задача состояла в том, чтобы узнать, с какими темами в сознании поэтов связывается образ Крыма. Помочь в решении этой задачи могло только тематическое моделирование, которое действительно открыло, что Крым — это прежде всего необычный пейзаж, сочетающий в себе летние и зимние элементы, своеобразная природа с морской доминантой и тема любви.

Примером работы тематического моделирования с художественной прозой могут быть статьи Т. Ю. Шерстиновой и соавторов[2].

Векторные модели

Одним из современных способов анализа текста является построение векторных моделей слов. Это технология, которая восходит к лингвистической идее дистрибутивной семантики (дистрибутивный от *distribution* — «распределение»).

Уже как минимум с 50-х годов XX века эта идея обсуждалась в фундаментальной науке. Она заключается в том, что слова получают свое значение только в контексте. Мы уже касались этой темы

[1] Лейбов Р. Г., Орехов Б. В. Между политикой и поэтикой: топика Крыма в современной русскоязычной наивной лирике // Шаги/Steps. Т. 8. 2022. № 2. С. 205–232.

[2] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic modelling with NMF vs. expert topic annotation: The case study of Russian fiction // Advances in computational intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Mexico City, Mexico, October 12–17, 2020: Proceedings / Ed. by L. Martínez-Villaseñor, O. Herrera-Alcántara, H. Ponce, F. A. Castro-Espinoza. Pt. 2. Cham: Springer, 2020. P. 134–151.

в связи с грамматической омонимией при лемматизации. Действительно, некоторые слова становятся понятными только в конкретной фразе, как форма «мыла» — существительное или глагол. Но это означает и обратное: контекст позволяет опознать семантику слова. Афористически это формулируется так: «Вы узнаете слово по той компании, в которой оно придет»[1].

Эта идея имеет несколько практических следствий. В частности, это должно означать, что слова, которые встречаются в похожих контекстах, имеют близкое значение. Вот два почти идентичных предложения: «я приду туда через несколько часов» и «я приду туда через несколько минут». Отличаются они только словами «час» и «минута», остальной контекст одинаков. Значит, и слова «час» и «минута» должны быть похожи по значению. И это действительно так, речь идет о двух единицах измерения времени. Правда, значения эти не идентичны.

Как мы можем рассчитать контексты слов, присутствующих в собрании текстов, представляющих для нас исследовательский интерес? Для таких вычислений строится таблица, в которой прописывается частота совместной встречаемости слов, то есть насколько часто слова попадают в один и тот же контекст.

Эти подсчеты позволяют представить контексты слов, а значит, и их семантику в виде вектора. Над векторами мы можем производить вычислительные операции — складывать их, вычитать один вектор из другого и т.д. Но главное — мы можем находить в векторном пространстве ближайшие векторы к данному. В практическом смысле это означает, что мы можем находить ближайшие по значению слова. При этом векторное пространство для слов многомерно, и число измерений зависит от числа слов, которые мы учитываем при анализе контекста.

Таким образом, представив семантику слова в виде математического объекта, мы можем вывести из этого несколько полезных следствий. Например, мы можем находить слова, близкие по значению, так называемые квазисинонимы. Это не настоящие синонимы, не то, что лексикографы помещают в синонимические словари. Хотя и настоящие синонимы среди квазисинонимов тоже могут присутствовать. Но квазисинонимами могут быть и антонимы — у антонимов часто очень близкие контексты словоупотребления. Вспомним также пример про «час» и «минуту». Это тоже не синонимы в строгом смысле этого термина.

[1] Firth J. R. Papers in Linguistics 1934–1951. London: Oxford, 1957. P. 11.

Правда, срабатывает такой поиск хорошо только тогда, когда слово частотно и когда мы подсчитали контексты для интересующего нас слова на достаточно большом корпусе. Ну и для более корректной модели мы должны лемматизировать тексты, поскольку носителем значения слова является все-таки лемма, а не отдельная словоформа. Иными словами, слово «окнами» не противопоставлено по значению слову «окном».

Еще раз подчеркнем, что речь идет о похожих с точки зрения векторных пространств словах, но не тождественных по значению.

Чем эта технология может быть интересна при работе с текстами на естественном языке, которые становятся объектами исследования в цифровых гуманитарных науках? Если построить такие векторные модели на контекстах, формируемых в цикле рассказов Конан-Дойла о Шерлоке Холмсе, а затем визуализировать векторное пространство, то видно, что Холмс и Ватсон из первого рассказа «Этюд в багровых тонах» не похожи на Холмса и Ватсона из других произведений цикла. Не похожи — значит, что эти имена попадают в отличные контексты. Объяснение очевидно: автор в начале пути еще нечетко представлял себе жанровые рамки и образы персонажей[1].

Кроме того, если провести аналогичное исследование для персонажей классических романов XIX века, то векторы для имен протагонистов этих романов оказываются близки друг к другу. Это означает, что авторы подбирают для описания своих протагонистов очень похожие слова, формируя близкие контексты. Точно такую же плотную группу векторов составляют имена персонажей, которые являются романтическим интересом главного героя.

С помощью векторов можно проводить и исследования поэтических текстов. Создав векторную модель на заранее заготовленной большой текстовой коллекции, можно затем оценить семантическую связность слов в нескольких жанрах. Можно взять, во-первых, статьи из Википедии, во-вторых, случайно сгененированные тексты, в-третьих, поэзию и прозу, написанные настоящими авторами.

Связность нескольких слов оценивается автором исследования[2] как значение их попарного сходства с точки зрения близости в векторном пространстве. Если упростить: слова, близкие по значению, будут связными, далекие — не связными. Оказалось, что Википедия

[1] Grayson S. et al. Novel2Vec: Characterising 19th Century Fiction via Word Embeddings // AICS. 2016. С. 68–79.

[2] Herbelot A. The semantics of poetry: A distributional reading // Digital Scholarship in the Humanities. 2015. Т. 30. № 4. С. 516–531.

демонстрирует наиболее высокую семантическую связность текста, случайно сгенерированные произведения — наименьшую. А поэзия в этом ряду занимает срединное положение, демонстрируя определенную неожиданность появления слов в стихах.

На русском материале такие исследования тоже проводились[1]. Если построить две векторные модели: одну на прозаических русских текстах, а другую на стихотворных, а потом запросить у векторной модели для каждого слова его квазисинонимы, то некоторые слова в той и другой модели не будут иметь общих квазисинонимов. Это будет означать максимальную разницу в семантике слов внутри прозаического и поэтического словоупотребления. Иначе говоря, слова эти выглядят как похожие, но по реальной семантике будут очень сильно отличаться: земля, любовь, человек, час.

Квазисинонимы слова «земля» показывают, что в поэтическом контексте обычно актуализируются символические значения, касающиеся погребального обряда (отпевание, погост, привал, территория), а в прозаическом — конкретные, связанные с сельскохозяйственной деятельностью (грунт, семя, перегной, пашня, почва). Так поэзия и проза актуализируют разные смыслы, а векторные модели позволяют это установить.

Стилеметрия

Стилеметрия (или *стилометрия* — сейчас под влиянием англ. *stylometry*) — это субдисциплина цифровых гуманитарных наук, переводящая стилевое своеобразие текста в исчислимые параметры, позволяющие измерять и количественно сравнивать стиль разных авторов.

Проблематичным при этом является и само понятие стиля, и набор параметров, и способ их квантификации.

Под стилем в контексте цифровых исследований неявно подразумевается неповторимая авторская манера письма, являющаяся текстовой репрезентацией личности пишущего. Иными словами, каждый человек уникален благодаря своему особому опыту, взглядам, навыкам. Эта уникальность, по мнению тех, кто использует

[1] Орехов Б.В. Стихи и проза через призму дистрибутивной семантики // Острова любви БорФеда: сборник к 90-летию Бориса Федоровича Егорова. СПб.: Росток, 2016. С. 652–655.

стилеметрию, должна проявляться и в сочиняемых им текстах, причем не на уровне тематики, а на уровне собственно языковой материи: выбора слов, использования грамматики, устойчивых сочетаний. То есть стилистическая особенность человека не в том, что он, будучи ихтиологом, пишет про рыб, или, будучи кинорежиссером, пишет о кино. Особенность в том, что один чаще, чем другие, употребляет слово «либо», а другой — оборот «на самом деле». В литературе о стилеметрии такая «особенность» называется авторским сигналом.

Такие стилистические особенности порой заметны невооруженным глазом и позволяют отгадывать авторство текста или адресата пародии.

Например, для речевой манеры И. Бродского характерно использование слова «суть» в качестве глагола-связки, равнозначной «есть»:

> Призрак бродит бесцельно по Каунасу. Он
> *суть* твое прибавление к воздуху мысли
> обо мне,
> *суть* пространство в квадрате, а не
> энергичная проповедь лучших времен.

«Суть» действительно является устаревшей формой глагола «быть» в 3 л. мн.ч., но Бродский, противореча языковым правилам, использует ее и для ед.ч., как в приведенном выше примере.

Этот характерный параметр помогает узнать авторскую манеру Бродского, но бесполезен для абсолютного большинства остальных авторов. Какие параметры в таком случае представляются продуктивными для стилеметрии?

Идейный прорыв в этом направлении произошел в последней трети XIX века. Он был связан с потребностями атрибуции, то есть установления авторства — и живописных полотен, и текстов. В области живописи идею, сходную с той, которая потом окажется высказана и для вербальных произведений, сформулировал итальянский искусствовед Джованни Морелли: «необходимо научиться отличать подлинники от копий. Однако для этого, утверждал Морелли, не следует брать за основу, как это обычно делается, наиболее броские и потому воспроизводимые в первую очередь особенности полотен: устремленные к небу глаза персонажей Перуджино, улыбку персонажей Леонардо и так далее. Следует, наоборот, изучать самые второстепенные детали, наименее затронутые влиянием той школы,

к которой художник принадлежал: мочки ушей, ногти, форму пальцев рук и ног»[1]. То есть наиболее репрезентативными становятся детали, которые меньше всего контролируются волей художника, те, которые он рисует машинально, по привычке. Такое внимание к деталям, перенос на них фокус исследовательского внимания культуролог К. Гинзбург называет «уликовой парадигмой», сравнивая деталь Морелли с уликой в криминальном расследовании.

Не случайно, что примерно в это же время (то есть в том же интеллектуальном контексте) идеи, очень сходные с основными постулатами «уликовой парадигмы», высказывает Томас Менденхолл[2]. Для установления авторства текстов, приписываемых Шекспиру, американский ученый предлагает использовать такой параметр, как длина слова в тексте. Достоинство этой характеристики опять-таки в ее неконтролируемости: в самом деле, вряд ли можно в обычной ситуации (то есть не включая сюда ситуацию авангардного творчества в духе литературы формальных ограничений[3]) представить себе автора, который возьмется выстраивать свой текст, высчитывая, сколько в нем присутствует слов той или иной длины.

Вторым важным обстоятельством является то, что длину слова и соответствующее ему количество в тексте мы можем перевести в числовые параметры и сравнить для нескольких произведений. Так, сопоставляя проблемный текст Шекспира с показателями Ф. Бекона и К. Марло, Менденхолл приходит к выводу о полном соответствии числовых показателей Шекспира и Марло.

Идея использовать длину слова для установления авторства себя не оправдала, то есть была позднее экспериментально опровергнута, но стремление ориентироваться на неконтролируемые автором параметры текста оказалось правильным. Вокруг такого рода «улик» и строилась в дальнейшем стилеметрия.

Вместо длины слова в качестве подсчитываемого параметра, отождествляемого со стилем, исследователи пробовали разные варианты:

– частотности слов и словоформ;
– цепочки символов;

[1] Гинзбург К. Мифы — эмблемы — приметы: Морфология и история. Сборник статей. М.: Новое издательство, 2004. С. 190.

[2] Mendenhall T.C. The characteristic curves of composition // Science. 1887. № 214s. P. 237–246.

[3] Бонч-Осмоловская Т.Б. Введение в литературу формальных ограничений. Литература формы и игры от античности до наших дней. Самара: Издательский дом «Бахрах-М», 2009.

– частотность и распределение частей речи;
– частотность грамматических конструкций;
– стиховедческие параметры (для поэтических текстов);
– длины слов и предложений;
– знаки препинания (зависит от уровня редакторского вмешательства в текст, может отражать представление о расстановке знаков препинания публикатором, а не автором).

Но и сам подсчет этих параметров может проводиться поразному. Например, одна из самых известных отечественных методик стилеметрического определения авторства («авторский инвариант» Т.Г. и В.П. Фоменко) предполагала вычисление процента служебных слов от числа всех слов в тексте. Но то же самое присутствие служебных слов в тексте может подсчитываться и иначе, например, с усреднением на определенный текстовый отрывок, скажем, в 10 тыс. слов.

Существенно, что серьезная проверка работоспособности этого набора параметров стала возможна только в последние десятилетия благодаря современному уровню оцифрованности текстов и компьютерным технологиям. В XX веке тестирование возможностей стилеметрии было вынужденно ограниченным: вручную проверялись сравнительно небольшие выборки текстов.

Еще одной проблемой стилеметрии стало то, что методики в ее рамках часто создавались под конкретные задачи определения авторства того или иного текста. С одной стороны, широко известные проблемы авторства культурно значимых произведений (диалоги Платона, трагедии Шекспира, «Конек-горбунок» Ершова, «Тихий Дон» Шолохова, спорные научные тексты Бахтина) подстегивали интерес исследователей к стилеметрическим проблемам. С другой стороны, многие методики создавались ad hoc, то есть для специального случая, и имели целью, таким образом, подкрепить точку зрения автора на конкретную историческую проблему. Работает ли та же самая методика для другого случая, оставалось второстепенным вопросом.

Выгодно выделяющимся на этом фоне универсальным (то есть не привязанным к конкретной историко-литературной проблеме) способом определения авторства стала Delta Берроуза, статья о котором была опубликована в 2002 году[1]. Дж. Берроуз предложил действовать следующим образом. Для исследования нужно собрать

[1] Burrows J.F. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship // Literary and Linguistic Computing 2002. 17(3): 267–287.

корпус текстов, в число которых входил бы текст, авторство которого было бы под вопросом, тексты, которые уверенно атрибутируются авторам, подозреваемым в том, что это они написали проблемный текст, а также некоторое количество текстов «фоновых» авторов той же эпохи. Заметим, что распространенной ошибкой многих стилеметрических исследований является привлечение произведений другой эпохи[1]. Стилеметрия чувствительна к изменению языка даже на коротких временных дистанциях. Берроуз поступил исследовательски чрезвычайно грамотно. Он выбрал текст, в авторстве которого никто не сомневается: «Потерянный Рай» Джона Мильтона. Правда, забавный факт в том, что в XVIII веке чрезвычайно недружелюбно настроенный к Мильтону интеллектуал Уильям Лоудер с помощью подлога пытался доказать, что «Потерянный рай» — это плагиат. Лоудер обвинял Мильтона в намеренных заимствованиях из сочинений авторов, писавших на латинском языке[2]. Другими текстами в исследовательском корпусе Берроуза стали другие произведения Мильтона и поэзия его современников.

Методика предполагала, что для расчетов задается некоторое количество наиболее частотных словоформ. Например, 100. Все дальнейшие операции производятся только для этих слов. Самыми частотными всегда оказываются служебные, а не полнозначные слова (см. об этом выше). Следствием этого обстоятельства является то, что Delta работает не с тематикой текста, которая выражается именами и глаголами, а с теми самыми неконтролируемыми параметрами.

Далее для каждого из выбранных слов в каждом из текстов корпуса вычисляется z-score — отношение разницы, взятой в процентах, от общего числа слов в тексте частотности слова в данном тексте и общей частотности слова по всему корпусу (то есть вычисленной для всех текстов выборки сразу, как если бы они вместе составляли один текст) к стандартному отклонению частотности слова по корпусу. Среднее арифметическое взятых по модулю разниц между z-score у двух сравниваемых текстов — это и есть значение стилистического расстояния между ними.

Эмпирически установлено, что это значение оказывается меньше у текстов, принадлежащих одному автору, и больше у текстов, принадлежащих разным авторам. В этом состоит основание для

[1] Например, Marina Iosifyan, Igor Vlasov And Quiet Flows the Don: the Sholokhov-Kryukov authorship debate // Digital Scholarship in the Humanities, Volume 35, Issue 2, June 2020, Pages 307–318.

[2] Уайтхед Дж. Серьезные забавы. М.: Книга, 1986. С. 112.

атрибуции. Действительно, подсчеты Берроуза подтвердили, что Мильтон — самый вероятный кандидат для авторства «Потерянного Рая», то есть методика прошла проверку.

Позднее выяснилось, что у метода, в целом достаточно надежного и работоспособного для разных языков, есть ряд ограничений.

Во-первых, сравниваемые тексты должны быть жанрово однородными. Сопоставлять вперемежку художественную, дневниковую, деловую прозу, стихи и драму с помощью Delta нецелесообразно, она не показывает осмысленных результатов.

Во-вторых, для надежного исследования нужны тексты не меньше 5 000 слов (а лучше 10 000 слов) каждый. На меньшем материале стилистически значимые статистики слов показывают случайные значения, поскольку не успевают стабилизироваться. Это, кстати, означает, что никакая методика определения авторства, основанная на статистике, не будет показательной для текстов небольшого объема.

Из-за недостаточности текстового материала исследователи, решающие задачи атрибуции, порой добавляют к ставшим традиционными подсчетам с помощью Delta дополнительные параметры, в частности, стиховедческие. Так, известный в филологической среде случай текстов, приписываемых поэту-декабристу Г. Батенькову, разбирается современными учеными с помощью добавления в анализируемые данные сведений о метрическом оформлении стихотворений[1]. В то же время благодаря масштабному исследованию, которое провел на материале поэтического корпуса НКРЯ Борис Орехов, известно, что стиховедческие параметры редко имеют в творчестве автора какой-то устойчивый тренд[2].

С начала 2000-х годов Delta (а также ее модификации вроде Delta Эдера) стала своего рода стандартом области цифровых гуманитарных исследований. Ее популярности способствовала не только вновь и вновь подтверждаемая надежность, но и низкий порог входа, обесепеченный программным пакетом Stylo для языка R[3].

---

[1] Шеля А., Плехач П., Зеленков Ю. Феномен Батенькова и проблема верификации авторства: многомерный статистический подход к нерешенному вопросу // Acta Slavica Estonica XI. Пушкинские чтения в Тарту. 2020. № 6. С. 131–165.

[2] Орехов Б. В. Микродиахрония стиховедческих параметров у русских поэтов // ВАПросы языкознания: Мегасборник наностатей. Сборник статей к юбилею В. А. Плунгяна / ред. А. А. Кибрик, Кс. П. Семенова, Д. В. Сичинава, С. Г. Татевосов, А. Ю. Урманчиева. М.: Буки Веди, 2020. С. 161–164.

[3] Eder M., Rybicki J., Kestemont M. Stylometry with R: a package for computational text analysis // R Journal. 2016. 8(1): 107–121.

Значимыми достоинствами этого пакета стали бесплатный характер его распространения и наличие графического интерфейса (а не только интерфейса командной строки, который создает психологические сложности для гуманитариев).

Для того чтобы воспользоваться этим программным пакетом, нужно установить на компьютер интерпретатор языка R (можно добавить к нему дружественную к пользователю R Studio), в командной строке установить пакет stylo:

```
install.packages("stylo")
```

Сделать это достаточно один раз.

Тексты для исследования имеет смысл подготовить следующим образом:

– каждый текст поместить в отдельный файл в формате plain text (с расширением.txt);

– все тексты должны быть в одной кодировке (предпочтительно: UTF-8);

– название файлов стоит выдержать в шаблоне: ИмяАвтора_НазваниеТекста.txt, то есть имя автора и название разделить знаком подчеркивания, для одного автора имя следует написать единообразно, это поможет при визуализации результата;

– не следует допускать пробелов и специальных символов внутри имени файлов;

– все файлы исследовательского корпуса нужно поместить в директорию с именем corpus.

После того как эта подготовка будет завершена, в интерпретаторе R нужно выполнить код, привязывающий программный пакет stylo к сессии:

```
library(stylo)
```

Этот код, в отличие от операции установки пакета (выше), нужно выполнять после каждого перезапуска интерпретатора.

Следующим шагом нужно установить рабочую директорию. Рабочей должна стать директория на уровень выше, чем директория corpus с файлами исследовательского корпуса, то есть та папка, в которой лежит папка corpus:

```
setwd('тут должен быть путь к директории, содержащей директорию corpus')
```

Далее нужно вызвать нужную функцию и работать уже с графическим интерфейсом:

```
stylo()
```

Stylo способен рассчитывать значение Delta для пар текстов в исследовательском корпусе и представлять результаты в виде таблицы попарных расстояний. Если текстов много, такая таблица будет сложна для визуального восприятия и анализа. Поэтому разработчики в качестве результата по умолчанию выбрали визуализацию в виде дендрограммы кластеризации, на которой тексты являются листами, а мера их стилистического сходства определяется дальностью или близостью ветвей. Наиболее близкие стилистически тексты (скорее всего, принадлежащие перу одного автора) в норме должны находиться на соседних ветках. Если выдержать шаблон именования файлов (см. выше), то тексты, заявленные как тексты одного автора (содержащие один и тот же префикс до знака подчеркивания) в цветном варианте графика будут помечены одним цветом.

Первая публикация, содержащая результаты применения Delta на русском языке, состоялась в 2016 году[1]. Исследователи Д. Скоринкин и А. Бонч-Осмоловская подтвердили работоспособность метода для русскоязычного материала. С тех пор с помощью Delta применительно к русскому языку решено несколько научных проблем, самой известной из которых можно считать проблему авторства «Тихого Дона».

«Антишолоховская» версия происхождения текста «Тихого Дона» появилась еще в первой трети XX века и к настоящему моменту получила широкую популярность. В контексте наличия множества частных фактов, которые можно трактовать как подкрепляющие любую из конкурирующих позиций, количественная атрибуция является для этой проблемы модельным случаем. Гораздо более надежным аргументом могли бы быть документальные свидетельства, подтверждающие ту или иную точку зрения, но таких

[1] Скоринкин Д.А., Бонч-Осмоловская А.А. «Особые приметы» в речи художественных персонажей: количественный анализ диалогов в «Войне и мире» Л.Н. Толстого // Электронный научно-образовательный журнал «История». 2016. Т. 7. № 7 (51).

свидетельств не существует. Количественные исследования также не могут подвести черту под дискуссиями, зачастую инспирированными вненаучными интенциями. Применение же непроверенных и не заслуживших авторитета (и скорее всего просто не работающих) цифровых методик, как это случилось в практике коллектива зарубежных русистов[1], только дискредитирует сам по себе цифровой подход к атрибуции.

В этой ситуации особенно любопытны были бы результаты, которые можно было получить для проблемы авторства «Тихого Дона» с помощью надежной методики Delta. Такое исследование провели отечественные ученые Н. П. Великанова и Б. В. Орехов[2]. Проанализировав межтекстовые расстояния для отдельных томов «Тихого Дона», «Донских рассказов» Шолохова, текстов В. Севского и Ф. Крюкова как наиболее вероятных авторов романа, с точки зрения антишолоховедов, а также художественных произведений современников Шолохова, филологи пришли к выводам, что:

– все тома «Тихого Дона», вероятнее всего, написаны одним человеком (в разных работах антишолоховского направления это положение ставилось под сомнение);

– наиболее вероятным является то, что «Донские рассказы» и «Тихий Дон» написаны одним человеком (не все антишолоховеды готовы признать и за «Донскими рассказами» авторство Шолохова, так что здесь нужны аккуратные формулировки);

– ни Севский, ни Крюков не являются вероятными авторами романа «Тихий Дон»;

– нет разницы между текстологически корректной версией романа и опубликованной с ошибками: Delta оказалась нечувствительна к правке текста[3].

Стилеметрию иногда отождествляют с атрибуцией. Это неверно. Когда мы говорим о стилеметрии, речь идет именно об измерении текстовых параметров, которые могут быть сопоставлены со стилем.

---

[1] Хьетсо Г., Густавссон С., Бекман Б., Гил С. Кто написал «Тихий Дон»? (Проблема авторства «Тихого Дона») / пер. А. В. Ващенко, Н. С. Ноздриной. М.: Книга, 1989. 186 с.

[2] Великанова Н.П., Орехов Б. В. Цифровая текстология: атрибуция текста на примере романа М. А. Шолохова «Тихий Дон» // Мир Шолохова. Научно-просветительский общенациональный журнал. 2019. № 1. С. 70–82.

[3] Данные для воспроизведения этого исследования опубликованы в Репозитории открытых данных по русской литературе и фольклору: Орехов, Борис, 2020, «Стилеметрические данные “Тихого Дона” и современной ему прозы», https://doi.org/10.31860/openlit-2020.05-R001, Репозиторий открытых данных по русской литературе и фольклору, V1.

Такое измерение может иметь прикладное значение для задач определения авторства, но спектр его применения шире и включает не только прикладные, но и академические задачи, формулируемые вокруг сопоставления текстов. В этом смысле стилеметрия наглядно показывает статус количественных исследований вообще: цифровые исследования не гарантируют точности (и, соответственно, не обеспечивают истинности выводов в задачах атрибуции). Они позволяют сравнивать объекты изучения. Без стилеметрии мы, оставаясь в научном поле, не можем сказать, насколько стилистически близок «Тихий Дон» Крюкову или Шолохову, а благодаря стилеметрическому подходу эта близость получает числовое выражение.

Следует заметить, что значимость атрибуции в случае художественных текстов сильно преувеличена. Центральным для культуры является текст, а не его автор. Именно благодаря тексту у читателя появляется интерес к автору, а не наоборот. Концепция смерти автора[1] добавляет оснований считать проблему авторства второстепенной. Дискуссионным является и вопрос о том, насколько один художественный текст может помочь при медленном чтении и интерпретации другого: художественный мир в каждом новом произведении конструируется автором заново. Поэтому предоставление об одном стихотворении методологически проблемно при переносе на другое стихотворение. Так что установление авторства некоторого произведения, вызывающего сомнения в своей атрибуции, не обязательно имеет научную ценность.

В то же время нельзя отрицать, что публику за пределами академического сообщества вопросы авторства живо интересуют. Ученым же представляется, что атрибуция — это наиболее естественный круг задач в гуманитарной сфере, где возможно применение количественных методов, поскольку предполагает вычислимый и проверяемый результат.

Однако гораздо более важной для гуманитарной науки является проблема понимания текста, его внутреннего устройства и механизмов взаимодействия с другими текстами внутри поля культуры.

Особую ситуацию как будто должен представлять философский текст. Философские произведения репрезентируют цельную философскую систему, освещая ее с разных сторон. От того, принадлежат ли автору философской системы те или иные слова, вроде бы зависит полнота нашего представления об этой системе, то есть

[1] Барт Р. Избранные работы: Семиотика. Поэтика. М., 1994 С. 384–391.

как раз задача атрибуции тут должна быть связана с проблемой понимания.

Но и здесь вопросы авторства находятся в подчиненном положении, как видно из следующей цитаты: «Диалог “Феаг”, который входит в платоновский корпус, но очевидно не принадлежит Платону, играет важную роль в развитии учения о божестве Сократа. Если в подлинных платоновских диалогах божество упоминается достаточно коротко, то в “Феаге” ему посвящена заключительная часть диалога»[1]. Центральным оказывается не вопрос об авторстве, а о принадлежности школе, исповедующей одну философскую традицию.

Одним из возможных направлений применения стилеметрии за пределами задач определения авторства оказывается исследование переводов. Известен эффект, при котором авторский сигнал при переводе проявляет себя ярче, чем сигнал переводчика. Иными словами, стилеметрически переводы текстов одного автора, выполненные разными переводчиками, объединяются в один кластер, а переводы разных авторов, выполненные одним переводчиком, нет.

На русском и английском материале этот эффект показан в статье о стилеметрии Набокова[2]. Б. В. Орехов с помощью Delta проанализировал, насколько похожи русскоязычные романы Набокова и русские переводы его англоязычных романов. Методика позволила увидеть два разных противопоставленных друг другу кластера, и даже переведенный Набоковым самостоятельно роман «Лолита» оказался в кластере переводных текстов и не смешался с русскоязычными. Кроме того, исследованию подвергся выполненный Набоковым перевод «Героя нашего времени» на английский язык. В сопоставлении с оригинальными романами Набокова и другими переводами Лермонтова на английский место набоковского перевода выявилось довольно определенно — в одном кластере с остальными вариантами англоязычного Лермонтова. Даже писатель с такой индивидуализированной авторской манерой, как Набоков, превращаясь в переводчика, приглушает свой сигнал в тексте, подчиняя его авторскому сигналу.

---

[1] Беликов Г. С. Речи Максима Тирского, посвященные божеству Сократа, в литературном и философском контексте I–II вв. н.э.: дис. … канд. философ. наук. М., 2020. С. 68.

[2] Орехов Б. В. Текст и перевод Владимира Набокова через призму стилеметрии // Новый филологический вестник. 2021. № 3. С. 200–213.

Еще одно исследование перевода с помощью стилеметрических инструментов выполнено на материале последней по времени попытки передать на русском языке «Илиаду» Гомера[1]. А. И. Любжин продолжил и завершил неоконченный перевод XVIII века, принадлежащий Е. Кострову. Исследовательская задача Б. В. Орехова состояла в том, чтобы сравнить эти переводы и измерить лежащую между ними стилистическую дистанцию. Работа показала, что дистанция эта значительна, и современному переводчику не удалось мимикрировать под стилистику поэта XVIII века, что может положительно сказаться на восприятии этой версии русского Гомера для современных читателей.

Кроме переводов, предметом стилеметрического исследования может быть текстовое оформление речи различных персонажей, масок, псевдонимных авторов. Так, в рамках диссертационного исследования, посвященного роману «Война и мир», Д. А. Скоринкин выявил индивидуальную стилистику речи персонажей[2]. В отдельной статье Д. А. Скоринкин и Б. В. Орехов[3] обнаружили, что Delta оказалась чувствительна к разнице в стиле т.н. гетеронимов, то есть вымышленных «авторов» с глубоко проработанной — в отличие от простых псевдонимов — биографией, которых создавал для подписи своих текстов португальский поэт Фернандо Пессоа. Анализ результатов, которые демонстрирует Delta, показывает, что стилеметрически тексты гетеронимов Р. Рейша, А. де Кампуша и А. Каэйра должны быть интерпретированы как тексты других людей, а не самого Пессоа.

[1] Орехов Б.В. «Илиада» Е. И. Кострова и «Илиада» А. И. Любжина: стилеметрический аспект // Аристей. 2020. Т. XXI. С. 282–296.

[2] Скоринкин Д. А. Семантическая разметка художественных текстов для количественных исследований в филологии (на примере романа «Война и мир» Л. Н. Толстого): дис. … канд. филол. наук. НИУ ВШЭ. М., 2018.

[3] Skorinkin D., Orekhov B. Hacking stylometry with multiple voices: Imaginary writers can override authorial signal in Delta // Digital Scholarship in the Humanities. 2023. Volume 38, Issue 3. P. 1247–1266.

Глава 7

Геоинформационные системы: подходы, методики, данные

(*Е. С. Гришин*)

Теоретические основы геоинформационных систем

Обращение в гуманитарных науках к пространственно ориентированным данным создает запрос на наличие соответствующих инструментов для решения задач, сопряженных с пространственным анализом. В период, предшествующий появлению цифровых технологий, все подобные инструменты обеспечивались средствами классической картографии. Развитие технологий, которое привело к появлению цифровых карт, а затем — к возникновению и использованию геоинформационных систем (ГИС), существенно расширило возможности оперирования пространственными данными.

Перечислим **основные преимущества**, которые предоставляют геоинформационные системы по сравнению с использованием обычных карт (некоторые из них имеют очевидный характер, однако должны быть названы для более ясного представления о назначении ГИС в гуманитарных науках):

– *учет и хранение данных*; геоданные могут находиться на различных цифровых носителях, их копирование может осуществляться как в полном объеме, так и отдельными частями; организация геоданных предполагает гибкую систему обращения к ним: пользователя или редактора может интересовать ГИС-проект в полном составе или его отдельные слои, объекты, атрибуты объектов;

– прямое продолжение функции хранения — *обращение к данным*; могут быть реализованы различные режимы доступа к ним:

с возможностью редактирования или без нее, запрос на различные наборы слоев с определенными общими характеристиками; существенно облегчается поиск и отбор объектов;

– *совмещение данных*; материалы из различных проектов или дискретные данные могут быть совмещены в рамках нового проекта;

– *анализ данных и его автоматизация* средствами ГИС-редактора.

По большому счету, методические и технологические основы использования ГИС носят универсальный характер и в равной степени применимы как в отношении естественных наук, так и гуманитарных. Однако есть определенные особенности последних, которые требуют указать на специфику использования геоинформационных технологий в гуманитарных исследованиях:

– *разнородность данных*; если в естественнонаучных ГИС-проектах преобладают унифицированные, единообразные по структуре и формату исходные данные (например, каталоги скважин, данные мониторинга наблюдательной сети), то в гуманитарных исследованиях часто приходится привлекать материалы, различные по своему происхождению и, соответственно, структуре. По этой причине автору приходится производить увязку исходных материалов, приводить их к общему виду;

– *степень достоверности данных*; при решении ряда научных задач в гуманитарных исследованиях приходится отражать уровень достоверности приводимых сведений, их предполагаемый или установленный характер. В естественнонаучных дисциплинах, связанных с ГИС, подобные проблемы также возникают, однако в другом виде и в меньшем объеме. В исторических геоинформационных системах сама локализация объектов может быть под вопросом и требовать дополнительных уточнений, а атрибутивные данные необходимо сопровождать указаниями на используемый источник;

– *необходимость структурирования данных*; при формировании данных приходится обращаться к неструктурированным источникам — нарративным текстам, множественным документам. Ситуация с ними заметно отличается от работы с готовыми статистическими материалами и каталогами объектов. Здесь задача редактора состоит в определении нужных атрибутов и заполнении их значений путем редуцирования текстовых материалов;

– *отсутствие готовых методических решений* при выполнении задач; во многом эта проблема упирается в ситуацию с методикой

картографирования: естественнонаучные картографические дисциплины традиционно хорошо обеспечены методической базой, которая используется и при работе с геоинформационными системами; между тем методика построения гуманитарно ориентированных карт и их цифровых аналогов в формате ГИС по многим вопросам остается слабо разработанной или вовсе отсутствует. В итоге приходится искать решения в каждом отдельном случае заново. Это дает некоторую свободу исследователю, но затрудняет накопление общей методической базы при проведении подобных работ.

Укажем на **назначение геоинформационных систем** в гуманитарных науках:

– *выполнение учебных задач*; в этом случае геоинформационный проект является инструментом для отработки навыков пространственного анализа, изучения тем с географической проблематикой; также ГИС-проект может выступать как своеобразный гид или навигатор по установленному набору объектов; во всех подобных вариантах главными качествами геоинформационной системы является ее доступность и наглядность, а также адаптированность по отношению к изучаемым навыкам или знаниям. При этом учебные ГИС-проекты не нацелены на получение новых результатов, они лишь выдают те значения и сведения, которые были заложены авторами для достижения целей обучения;

– *получение и обработка справочной информации*; подобно учебным проектам, этот класс геоинформационных систем ориентирован на хранение и вывод заложенной информации, однако имеет ряд отличий, в числе которых значительный объем данных, их более сложная структура, наличие справочных служебных таблиц и классификаторов; все это позволяет не просто получать статичную информацию по поиску, но и формировать сложные выборки; таким образом, при должном подходе хорошо проработанные справочные ГИС могут быть отчасти исследовательским инструментом, пусть и со значительными ограничениями, ведь они имеют пользовательский характер, без возможности ввода новой информации с клиентской стороны;

– *осуществление исследовательских задач*; здесь ГИС является инструментом исследования, средством получения новых данных, которые не были заложены в готовом виде в исходных материалах. Сюда же стоит добавить, что ГИС-проект сам по себе может являться основным результатом исследовательской работы наравне с текстом, инфографикой и другими результатами. Геоинформационные

проекты подобного рода агрегируют в себе множество функций, включая учебные и справочные, являясь вместе с тем мощным ресурсом для решения пользовательских задач.

Конечно, строгое разграничение перечисленных назначений не требуется, однако автору проекта стоит заранее определиться с основной целью построения геоинформационной системы, так как это во многом определит стратегию работы с ней и востребованность конечного результата.

Осталось обозначить, в чем состоит преимущество ГИС-редакторов по сравнению с обычными графическими редакторами при подготовке картографических материалов или решении частных пространственных задач. Почему, например, при создании макетов с инфографикой не использовать простой векторный редактор, в котором можно построить нужные объекты, оформить их и сделать соответствующие подписи? Этот вопрос справедлив в том смысле, что между ГИС-редакторами и редакторами векторной и растровой графики действительно есть некоторые общие функции, в частности, послойное представление данных. Также можно согласиться, что для решения частных оформительских задач функций векторного редактора может оказаться достаточно. Однако специализированные программы по работе с ГИС предоставляют достаточное количество возможностей, которые резко ускоряют работу с любыми пространственными данными. Прежде всего, это математический аппарат, который упрощает действия с проекцией и масштабом ГИС. В рамках одного проекта могут быть совмещены разнородные данные, взаимная увязка которых не вызывает затруднений, что было бы невозможно при использовании обычного векторного редактора. Кроме того, функционал ГИС автоматизирует многие операции по графической обработке атрибутивных данных — построению картограмм, картодиаграмм, типологической дифференциации объектов.

Если перед автором ГИС стоит задача не только получить некоторый исследовательский результат, но и представить свой проект как открытый цифровой ресурс, он может осуществить это через публикацию собственных геоданных на доступном картографическом сервисе. Другой вариант состоит в полноценной публикации авторского ГИС-проекта на веб-ресурсе с сохранением возможностей интерактивного взаимодействия с объектами и слоями.

В отношении времени, пространственно-хронологического анализа геоинформационные системы могут быть определены как

ретроспективные (исторические), если отражают состояние изучаемого пространства в прошлом; современные, отражающие актуальную информацию о регионе; мониторинговые, данные которых обновляются в реальном времени и отражают любые доступные для фиксирования изменения; динамические, аккумулирующие разновременные данные. Хронологическая ориентация ГИС характерна для многих исследовательских проектов и требует заметного усложнения их структуры.

Пространственная применимость и охват исследуемой территории характеризует **уровень ГИС**, который может быть определен следующим образом:

– *муниципальный уровень*; ГИС ориентирован на конкретную местность, населенный пункт, муниципальное образование; к ним могут относиться проекты по местной экологии, земельному кадастру, различные политематические цифровые ресурсы;

– *региональный уровень*; охватывают один или несколько регионов, при этом потенциальная пользовательская база может быть гораздо шире, чем собственно жители рассматриваемой территории; помимо уже названной экологической тематики, могут быть составлены в туристической направленности, а также как инструмент мониторинга состояния региона, планирования его развития;

– *национальный уровень*; все, относящееся к региональному уровню, будет справедливым и для уровня всего государства, со значительным расширением тематического применения: ГИС-проекты по анализу уровня образования населения, отслеживания демографических процессов, решение любых других пространственных задач национального значения;

– *международный и глобальный уровень*; характерен для картографических сервисов, охватывающих всю территорию Земли. Тематическое направление и уровень их проработки может сильно варьироваться от одного проекта к другому и не поддается строгой регламентации. Характерно комбинирование данных, создание аккумулирующих систем, в которых объединяются сведения из различных источников.

Если продолжать сравнение ГИС с картами классического вида, то территориальный уровень геоинформационных систем может быть сопоставлен с базовым масштабом карт. Так, муниципальные ГИС могут быть представлены как аналоги крупномасштабных топографических карт, национальные — как мелкомасштабные карты и т.д. Стоит, однако, дополнить, что подобно картам ГИС способны

между собой взаимодействовать и дополнять друг друга; так, глобальные ГИС не возникают внезапно, они формируются как сумма многочисленных проектов национального уровня; те в свою очередь формируются за счет сбора данных более низового уровня и т.д.

Классическое **определение** ГИС, состоящее в том, что это информационные системы, которые осуществляют сбор, хранение, обработку и визуализацию пространственных данных, в полной мере отражает основные задачи, возлагаемые на них. Соответствующим образом группируются и подсистемы в составе ГИС. По определению М. Н. Де Мерса, геоинформационная система состоит из четырех подсистем, которые осуществляют:

– сбор данных;

– хранение и выборки данных;

– манипуляции данными, их редактирование, анализ и другие операции;

– вывод данных в различных графических формах: картографических материалах.

Большая часть этих подсистем относится к функционалу программных средств, предназначенных для работы с геоинформационными системами.

Независимо от используемого программного обеспечения, автор ГИС-проекта будет обращаться к определенным **элементам управления**, без которых невозможна работа ни в режиме редактирования, ни в режиме чтения. Большая часть этих элементов носит типовый характер и присутствует с определенной вариативностью в большинстве ГИС-редакторов. Компоновка интерфейса также, как правило, носит сравнительно унифицированный вид и может быть описана как набор следующих блоков:

– *блок управления проектами*; по сути, это меню проектов, которое позволяет создавать проекты, сохранять в них изменения, добавлять новые данные, экспортировать макет в растровую карту и прочее;

– *блок «легенды»*, который представляет собой структурированный список задействованных в проекте данных: слои, объединенные в группы; это один из важнейших блоков, поскольку он фактически отражает содержательную структуру ГИС-проекта и позволяет обращаться к настройкам каждого отдельного слоя, менять порядок слоев при отображении на карте;

– *блок с различными функциональными меню*; к основным наборам инструментов для работы с ГИС-проектом относятся:

меню редактирования геометрии объектов; менеджер управления условными знаками и стилями геоинформационной системы; меню навигации и настроек отображения проекта (экстент, масштаб, режимы работы и прочее); кроме того, обычно в ГИС-редакторах присутствуют многочисленные панели инструментов для решения специализированных задач или более продвинутой работы с геометрией, растровыми данными, конвертацией и другими задачами. Значительная часть этих инструментов также присутствует в большинстве программных средств;

– *блок основного содержания*, который отображает графический вид рабочего пространства и объекты; это область, в которой автор видит результат своей работы, и одновременно взаимодействие с ней позволяет осуществлять обращение к отдельным объектам, пространственные выборки, перемещаться по картографируемой территории, строить новые объекты, редактировать геометрию уже существующих и выполнять другие операции;

– *блок служебных показателей проекта*; он, как правило, занимает нижнюю часть интерфейса ГИС-редактора и отображает текущую информацию по работе с проектом: координаты курсора в рамках рабочей области, текущий рабочий масштаб (в зависимости от программного средства);

– отдельно стоит выделить такой элемент интерфейса, как *атрибутивные таблицы*, которые относятся к отдельным слоям; они сопровождаются собственной панелью инструментов, которая позволяет делать сортировку объектов по отдельным полям, осуществлять выборку объектов по атрибутам, обращаться к функциям калькулятора полей, редактировать атрибутивные данные.

Структура геоинформационного проекта и используемые данные

Структурные единицы геоинформационной системы:

– *проект*; в своей цифровой версии проект представляет собой файл с расширением используемого ГИС-редактора и фактически является набором ссылок на различные ресурсы; в нем также содержится определенная информация: настройки самого проекта, структура данных и применяемые к ним правила отображения,

однако сведения об объектах, то есть пространственные и атрибутивные данные, в файле проекта не содержатся;

– *группа слоев*; слои могут быть сгруппированы по логическому, содержательному или пространственному признаку; например, слои по гидрографическим объектам; слои с линейной геометрией; группы слоев формируются непосредственно в проекте и отображается в блоке «легенды», то есть основного содержания проекта; объединение слоев в группы позволяет легче ими управлять (допустим, если необходимо отключить отображение всех слоев той или иной группы), также это облегчает восприятие структуры проекта; группы могут составлять сложные структуры и иметь вложенный характер;

– *слой*; набор объектов с общим типом геометрии, атрибутивной таблицей и уникальным заголовком; объекты в рамках одного слоя могут дифференцироваться по типологии, отображению на карте (некоторые из них могут вовсе не отображаться), однако их объединяет общий набор данных и структура применяемых параметров (атрибутов); если группы слоев носят лишь логический и структурный характер, то сам слой имеет четкое выражение в виде набора файлов или таблицы в базе геоданных;

– *группа объектов*; группы могут иметь постоянный или динамический характер; в первом случае объекты объединяются в группу по общим атрибутивным признакам, что может быть отражено путем присвоения им соответствующих условных знаков или подписей; в случае с динамической группой объекты попадают в ее состав в результате действий пользователя или редактора — путем осуществления атрибутивной выборки различной степени сложности.

Основные этапы работы над ГИС-проектом

Работа над ГИС-проектом состоит из нескольких этапов:

– определение задач проекта, его содержания и структуры;

– обработка источников, первичных материалов для использования в проекте;

– формирование цифровой модели ГИС-проекта, адаптация картографической основы;

– наполнение ГИС основным содержанием: векторизация объектов, заполнение атрибутов;

– выполнение исследовательских действий над полученными данными, если это предусматривается задачами проекта;

– оптимизация и рецензирование проекта, исправление выявленных ошибок; публикация на цифровых ресурсах.

Конечно, эти этапы в ходе прикладной работы могут быть уточнены. Так, при обработке источников может возникнуть потребность в пересмотре намеченной структуры, добавлении новых атрибутов.

В содержательной плоскости ГИС состоит из двух основных блоков: картографической основы и тематического содержания. Если тематическое наполнение составляет основную работу автора, то пути подготовки картографической основы имеют несколько конкретных вариантов.

Прежде всего нужно определиться, что входит в ее состав и что должно быть оставлено доступным для пользователя после адаптации основы. Во многом это зависит он основного содержания и целевого назначения ГИС-проекта. Так, в одном случае картографическая основа может быть лишь графическим фоном, который способствует общему ориентированию и навигации в пространстве; в другом — элементы основы являются важными составными частями содержания, обосновывают и дополняют его. Особенно это касается проектов, посвященных анализу пространственной активности, построению маршрутов, туризму и другим подобным темам.

К базовым частям картографической основы относятся гидрография, рельеф (часто может быть разгружен или отображен минимально), населенные пункты, пути сообщений, важнейшие объекты инфраструктуры, границы административного деления. Включение в состав основы проекта тех или иных блоков, а также их графическое отображение зависит от того, насколько эти действия будут соответствовать основным фактографическим слоям.

Что касается возможностей подготовки картографической основы, назовем следующие из них:

– подключение внешнего картографического сервиса вроде Open Street Map (OSM); этот вариант представляется самым простым и доступным по реализации, однако его использование существенно ограничивает возможности адаптации содержания: автор фактически вынужден использовать основу в том виде, в котором она представлена на сервисе;

– использование открытых геоданных в виде отдельных слоев; многие цифровые ресурсы и хранилища геоданных позволяют использовать свои материалы для подключения к авторским проектам; этот вариант представляется довольно удобным и сбалансированным, так как сохраняет возможности адаптации и редактирования основы, при этом ее сборка не представляется особо трудоемкой и затратной по времени;

– построение основы по первичным данным путем векторизации растровых карт, подключения космоснимков и других источников; наиболее требовательная к навыкам и времени автора процедура, однако она может оказаться единственно возможной, если тематика проекта выдвигает высокие требования к точности картографической основы; так, для многих исторических ГИС важна проработка топографии региона соответственно изучаемому хронологическому интервалу, что исключает использование готовых геоданных и внешних картографических сервисов.

Первичные материалы (источники) в зависимости от носителей, формы подачи информации и происхождения могут быть разделены на несколько основных типов:

– текстовые; к ним относятся нарративные (повествовательные) тексты, в которые интегрированы пространственные данные; описательные тексты, которые могут быть представлены как текстовая модель пространства; официальные документы, например тексты мирных договоров или постановления об изменении административного деления; разнообразные статистические справочники, перечни и другие упорядоченные, формализованные текстовые материалы;

– картографические; это один из важнейших источников при ретроспективном и тематическом картографировании, когда автору, помимо переработанного текста, нужно привлекать карты своих прямых предшественников или картографические материалы далекого прошлого;

– данные дистанционного зондирования Земли (ДЗЗ); к ним относятся космические и аэроснимики, а также результаты геофизического исследования земной поверхнности;

– источники особых категорий: иконографика, кинофотодокументы, памятники материальной культуры, топонимия региона и др.

Каждый тип источников требует собственной методики обработки и преобразования данных в адаптированный для геоинформационной системы вид. В этом отношении *нарративные источники* представляют особую сложность, так как не содержат

структурированную пространственную информацию: она распределена в тексте в виде отдельных вставок, экскурсов, упоминаний локаций. Соответственно, для преобразования нарративного текста в модель пространства в виде геоинформационной системы понадобится его редукция на отдельные элементы, которые поддаются геокодированию. Самый очевидный способ: составление перечная всех географических объектов в тексте для его дальнейшей обработки. Этот процесс аналогичен составлению географического указателя к тексту, хотя и подразумевают более детализированную информацию. Так, в ходе составления подобного перечня учитываются типы объектов, частота их упоминаний в тексте, формы топонимов, событийные поводы, связанные с их упоминаниями. Затем проводится локализация объектов из списка, внесение их в соответствующий слой ГИС и заполнение атрибутов из источника.

Статистические данные эффективно обрабатывать с помощью создания связанных электронных таблиц. Тем самым удается сохранить оригинальную структуру данных, которая была задана источником, и при этом избежать нагромождения полей в атрибутивной таблице.

Работа с *картографическими источниками* строится совершенно иначе, так как они могут прямо внедряться в состав ГИС-проекта. Растровые карты (сканы бумажных карт или цифровые версии макетов) «привязываются» к основе ГИС, другими словами, поверхность растра приводится в соответствие с картографической моделью поверхности Земли путем добавления определенного набора контрольных точек. Осуществление привязки происходит в ходе некоторых действий. Прежде всего, это оценка самой растровой карты — пригодна ли она для привязки, какова ее математическая основа, нуждается ли растр в предварительной обработке. Если проекция привязываемой карты отличается от проекции ГИС-проекта, последнюю рекомендуется привести в соответствие с источником, чтобы привязка прошла корректно. Затем карта добавляется в состав проекта, и автор может начать добавление контрольных точек. Если привязываемая карта имеет современный вид, то наиболее эффективным вариантом будет установление контрольных точек по координатной сетке. Предпочтительно их равномерное распределение по углам и периферии растра. В другом случае контрольные точки могут быть определены по стабильным географическим объектам, которые присутствуют и на картографической основе, и на растровой карте. В дальнейшем можно использовать различные

алгоритмы трансформации привязываемой карты, что позволит избежать искажений и несоответствий.

Стоит, однако, оговориться, что в ряде ситуаций привязка растровой карты оказывается ненужной. Так, сканированная карта может быть источником актуализации отдельных данных, для этого ее привязка не обязательна. Кроме того, многие исторические карты и вовсе не могут быть эффективно задействованы в привязанном виде.

Для *данных ДЗЗ* в функционале ГИС предусмотрены особые инструменты по обработке снимков территории, автоматическому дешифрированию объектов, обработке тех или иных каналов, калькуляции растров и прочему. Дешифрирование космо- и аэроснимков составляет особое направление в геоинформатике и требует глубокого знания темы, связанной с задачами дешифрирования. Стоит отметить при этом, что данные ДЗЗ чаще используются в естественнонаучном картографировании, однако в отдельных темах могут быть задействованы в ГИС гуманитарной направленности.

Математическая основа ГИС-проекта состоит из базового масштаба, проекции и используемой системы координат.

Выбор *масштаба* обусловливается уровнем самой ГИС (как уже отмечалось, для муниципального уровня актуальным будет крупный масштаб, для национального и глобального — мелкий); детальностью и насыщенностью используемых данных — чем больше плотность объектов на изучаемой территории, тем более будет обоснован выбор в пользу большего масштаба; используемыми источниками: если в основе проекта применяются картографические источники среднего масштаба, например, 10-верстовые карты, то базовый масштаб конечного результата также будет соответствовать этому показателю. Конечно, на пользовательском уровне масштаб не является строгим показателем и может быть смещен в ту или иную сторону за счет зумирования, однако его изначальное определение все же необходимо для самой работы над проектом. Выбор масштаба влияет на подход к работе с картографической основой, отображение объектов. Так, те объекты, которые в крупном и среднем масштабе имели площадной вид, на мелком отображаются как точечные знаки.

Отметим также практику построения полимасштабных проектов, когда по мере выставления того или иного уровня масштаба меняется и отображение основного содержания. Стоит иметь ввиду, что эффективное применение полимасштабного подхода требует проработки большого объема данных, которые обеспечивают каждый

масштабный уровень своим набором объектов. Это практически равносильно созданию нескольких карт одной территории с разной степенью детальности.

Что касается используемой проекции, ее прикладная роль для автора ГИС важна в двух аспектах: отображение картографического содержания ГИС для пользователя и манипуляции с проекцией в ходе привязки картографического источника, о чем уже шла речь выше. В ГИС-редакторе в настройках проекта предусматривается возможность изменения различных параметров проекции и системы координат, при этом автор может использовать готовые решения, доступные в самой программе.

При создании нового слоя определяется также его геометрический тип, среди которых выделяются три основных:

– *точечные объекты*; их геометрия характеризуется только координатной парой; точечные объекты применяются для обозначения тех объектов, линейными и площадными размерами которых можно пренебречь;

– *линейные объекты*; могут иметь замкнутый (как в случае с полилиниями) или открытый характер; также могут иметь направленный вид, если используются как указатели движения маршрутные линии;

– *полигональные объекты.*

Обработка геометрии предполагает выполнение ряда действий с уже имеющимися объектами, помимо редактирования их вершин или добавления новых участков: объединение/слияние нескольких объектов в один; разделение объекта на отдельные части по заданной точке или секущей линии; огрубление геометрии объектов (для генерализации) или сглаживание угловатых участков; поиск точек пересечения объектов; оверлейный анализ (совместная обработка наложения двух или более исходных слоев одной географической области, в результате которой создается производный слой с новыми географическими данными как комбинация топологических сегментов исходных географических данных); построение буферных зон от заданных объектов; поиск объектов, входящих в ареалы объектов других слоев. Все эти операции позволяют получить новые результаты, недоступные напрямую в готовом виде из первичных материалов.

Работы с атрибутивными данными. *Атрибуты* — наборы параметров объекта, которые могут быть задействованы для его определения или изучения средствами ГИС. Совокупность всех

полей атрибутов и их значений в рамках целого слоя представляет собой *атрибутивную таблицу*. Содержательно атрибуты могут быть разделены на следующие группы:

– *описательные* (качественные); качественные характеристики носят компактный характер, однако при этом содержат информацию, которая не может быть в полной мере унифицирована и упорядочена; часто атрибуты подобного рода используются для вспомогательных комментариев, коротких описаний объектов, которые не укладываются в формат остальных атрибутов; стоит обозначить, что неформализованный характер этих атрибутов и текстовом формате затрудняет их использование при поиске, осуществлении выборки объектов;

– *типологические*; имеют ограниченный набор возможных значений и, как правило, сопровождаются типологическим справочником, в котором указываются индексы типов, их значение и расшифровка; именно типологические атрибуты используются для формирования отдельных групп объектов в рамках одного слоя, а также для оформления объектов различными условными обозначениями;

– *числовые*; применяются для фиксирования разнообразных числовых показателей: дата появления объекта, численность населения, площадь и проч.; числовые показатели имеют уникальные значения, их формат позволяет построение картограмм и картодиаграмм.

Также можно выделить отдельные разновидности служебных атрибутов, которые могут быть предназначены для идентификации объектов, их связи с другими таблицами или слоями.

Назначение атрибутов имеет не только сопроводительный, справочный характер, они выполняют ряд важнейших функций:

– *идентификация объектов*; преимущественно это осуществляется с помощью числового атрибута id, который уникально задается для каждого отдельного объекта; однако для многих слоев актуальным будет также использование текстовых номинативных атрибутов (названий); это не отменяет, конечно, возможности совпадения названий нескольких разных объектов, однако подобная ситуация сама по себе может быть темой пространственного анализа, например, при изучении топонимов того или иного региона;

– *вывод подписей к объектам*; подобно классическим картам, геоинформационные системы используют текстовую и символьную семантику для дополнительной информативности своей графики, и подписи в этой связи играют едва ли меньшую роль, чем

графические объекты; чаще всего в качестве подписей выносятся те же атрибуты, что и для идентификации, однако для решения конкретных исследовательских задач для отображения подписей могут быть задействованы любые другие доступные атрибуты; кроме того, в ГИС-редакторах предусмотрен механизм составного подписывания объектов с помощью скриптовых языков, что позволяет выводить на карту несколько атрибутов с дополнительной синтаксической обработкой;

– *визуализация объектов с помощью условных обозначений*; если автор предполагает показать различную типологию объектов на цифровой карте, то атрибуты могут быть применены как механизм их дифференциации с использованием различных условных знаков для каждого отдельного типа; если же в качестве определяющего атрибута выбирается числовой, то графическое отображение значений атрибутов также может быть применено путем различных цветовых заливок или с помощью размеров условных знаков;

– *поиск и выборка объектов*; таблица атрибутов сопровождается формой для осуществления атрибутивной выборки, в которой пользователь должен выбрать нужные поля, применить нужные условия и указать значения; как и в SQL-запросах, могут использоваться различные математические и логические выражения и операторы; результатом поиска становится выделение объекта и в формате таблицы, и на карте;

– *получение значений статистики по слою*; это может быть осуществлено как в ходе выборки, так и при обращении к всем объектам слоя; так, пользователь может уточнить общее количество объектов в слое, максимальные, минимальные и средние значения того или иного поля и другие обобщающие значения;

– *получение картометрических показателей по объектам*; речь идет о пространственных параметрах, которые могли быть неизвестны заранее, по используемым источникам, и становятся доступны через обращение непосредственно к функциям ГИС; обычное название звучит как «вычисление геометрии» и позволяет установить площадь и периметр для полигональных объектов, протяженность для полилиний, координаты для точечных объектов, а также другие доступные характеристики (например, координаты центроида);

– *проведение анализа и получение новых сведений* с помощью функций «калькулятора полей».

Типы данных в атрибутивных таблицах аналогичны форматам данных в электронных таблицах и базах данных:

– текстовое поле с установленным количеством символов;
– целое число;
– число с плавающей запятой;
– дата.

Выбор типа данных определяет дальнейшие возможности по работе атрибутами конкретного поля. Так, выбор текстового формата для поля с датами приведет к неправильной сортировке записей и не позволит выдержать унификацию данных. Точно так же к разнородности данных может привести использование текстового формата для типологических атрибутов.

Калькулятор поля и аналогичные функциональные инструменты ГИС-редактора производят присвоение значений атрибутов сразу всем объектам слоя или только тем из них, которые попали в результат выборки. Значение атрибута устанавливается на основе равенства, более сложного выражения или выполнения определенного скрипта. Так, в ходе применения калькулятора полей может быть осуществлено вычисление геометрических показателей или манипуляции с ними (например, установление площади полигонов и вычисление плотности населения по ним), присвоение значений полей на основе модификации исходных данных; копирование значений из одного поля в другое с изменением формата данных; массовое присвоение новых типологических значений выбранным объектам. Все подобные операции заметно ускоряют как наполнение содержания ГИС-проекта (иначе пришлось бы вручную заполнять те значения, ввод которых производится автоматически с помощью калькулятора полей), так и проведение аналитических действий над материалом.

Пространственный анализ средствами геоинформационных систем

Как уже говорилось, для использования ГИС в проведении исследований необязательно построение собственных проектов. Часто бывает достаточно использования уже имеющихся ГИС, особенно для решения частных задач вроде поиска и локализации объекта,

получения картометрических и морфометрических показателей, осуществления выборки.

Другое дело, если полученные данные требуют дополнительной интерпретации и их преобразования в новые пространственные объекты, отдельные слои со своей атрибутивной нагрузкой.

Для графической визуализации собственных данных автору необходимо выбрать те способы отображения, которые соответствуют геометрическому типу объектов, отображаемым показателям и применяемой методике пространственного анализа. Для типологического отображения точечных знаков используется *значковый способ*: каждому типу объектов присваивается собственный знак, при этом может использоваться как простая дифференциация (когда за типологию отвечает один атрибут), так и сложная; в последнем случае разделение объектов может производиться по двум или трем атрибутам одновременно. В этой ситуации за графические отличия будут отвечать различные элементы знака: за базовый тип — определенный геометрический символ, за дополнительные характеристики и подтипы — цвет заливки знака, обводка или дополнительные геометрические элементы.

Если же основная атрибутивная нагрузка на точечные объекты представлена числовыми показателями, то применяется способ *круговых картограмм*: каждый объект визуализируется в виде круга, разделенного на секторы; радиус круга определяется общей суммой всех задействованных показателей, а секторы отображают долю тех или иных полей от общего числа.

Линейные объекты обладают большим потенциалом в плане разнообразия графической интерпретации данных. Так, типология может быть отражена как с помощью цвета линии, так и через паттерн ее отображения (штриховой, пунктирный, штрихпунктирный). Числовые показатели визуализируются путем утолщения линий пропорционально значениям. Если линия ограничивает территорию какого-либо явления или процесса, имеющего направленный характер, то в ней могут быть задействованы бергштрихи, показывающие направление. При визуализации изометрических процессов в разрыве линии приводятся значения показателей. Оба последних способа применяются, например, при построении эпидемиологических карт.

Площадные объекты объединяют в себе наибольший набор средств графического отображения различных показателей. Для различных по своим характеристикам территорий применяются способы выделения *ареалов и использования качественного фона*:

полигоны заливаются цветом, определенным паттерном площадной штриховки или крапом (набором распределенных по площади маркеров). При этом в особо насыщенных данными проектах могут совмещаться сразу несколько вариантов — и заливка цветом, и штриховка с крапом.

Не менее разнообразны и способы визуализации числовых показателей по площади. Самый простой из них — способ количественного фона, когда каждому диапазону значений соответствует свой цвет или другой вид заливки полигона. Для более сложных случаев, когда необходимо отобразить множество показателей, применяются картодиаграммы, а также точечный способ отображения площадной статистики (1 точка соответствует 1 единице показателя на площадь).

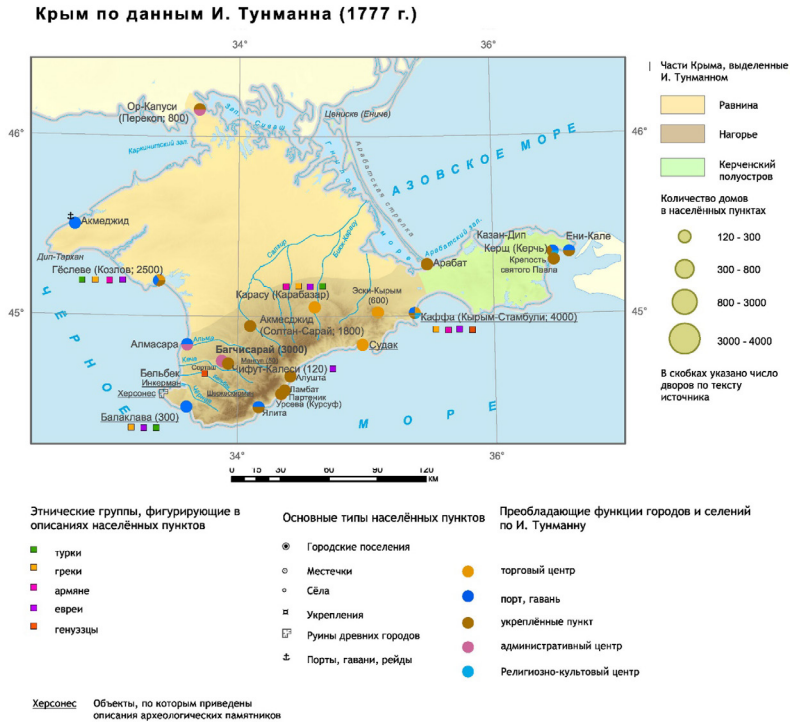


Рис. 7.1. Карта Крыма по описанию И. Тунманна. Пример преобразования текстового описания региона в пространственную модель с сохранением атрибуции и типологии объектов по используемому источнику

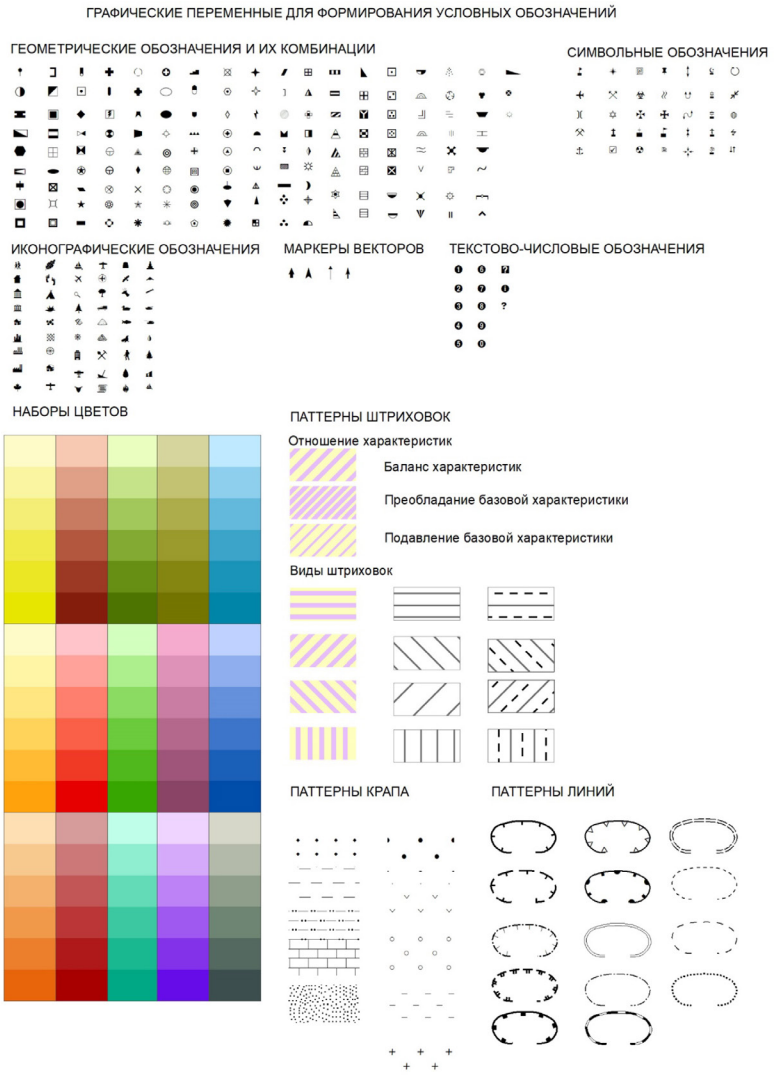


Рис. 7.2. Наборы графических переменных для формирования условных обозначений для отображения различных показателей объектов: типологической принадлежности, качественных характеристик и числовых показателей

Рис. 7.3. Карта Реконкисты демонстрирует типичное использование цветовых изохронов: каждый цветовой ареал отображает территорию с общим хронологическим показателем. Этим же способом отображаются прочие динамические процессы, имеющие характер распространения из одного ареала

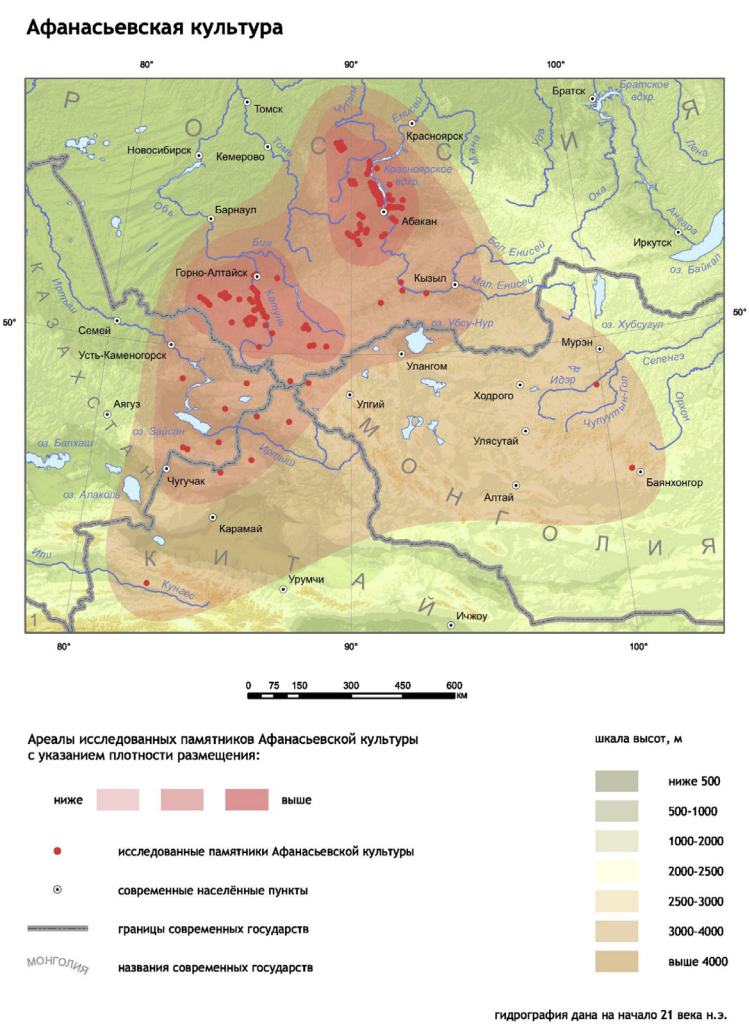


Рис. 7.4. На карте афанасьевской культуры отображены ареалы плотности распространения памятников, которые были выделены с помощью инструментов ГИС-редактор путем построения сетки и установления плотности объектов в каждой ячейке

- I – картографирование территории Тамбовского уезда на межевых планах (к. XVIII в.)
- II – съемка силами межевого ведомства и военных топографов (1850-е гг.)
- III – топографическая съемка Генштаба РККА (1930-е – 1941 гг.; лист N-37-131 масштаба 1 : 100 000)
- IV – современная топографическая съемка местности по тому же листу на 1988 г. с последующей актуализацией.

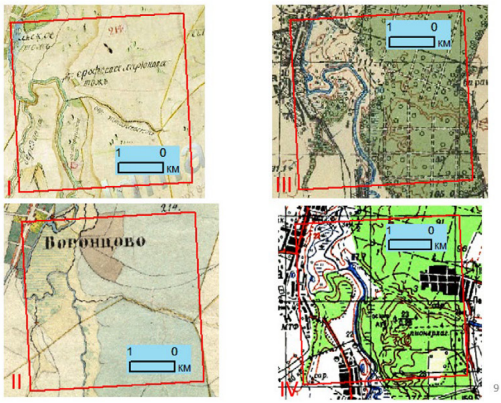


Рис. 7.5. Показаны разновременные картографические источники одной и той же местности с хронологическим охватом с конца XVIII до конца XX в. Сопоставление такого рода дает возможность проследить развитие территории, ее освоение и даже динамику топографических условий; однако вместе с тем стоит учитывать, что карты различных периодов строились по своим методикам, которые могли отличаться в наборе картографируемых объектов, точности их локализации, используемой семантики

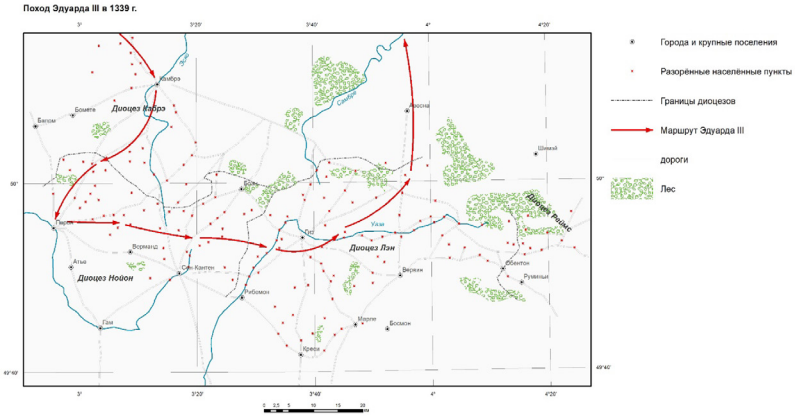


Рис. 7.6. На карте похода Эдуарда III в 1339 году в ходе Столетней войны показаны разоренные населенные пункты. Они образуют своеобразную буферную зону, прилегающую к маршруту похода. Подобные буферные зоны могут быть установлены эмпирически и моделироваться в ходе дальнейших исследований. Особенно эффективно построение буферных зон при изучении явлений и процессов, имеющих очаговый характер распространения: природные бедствия, пожары, эпидемии

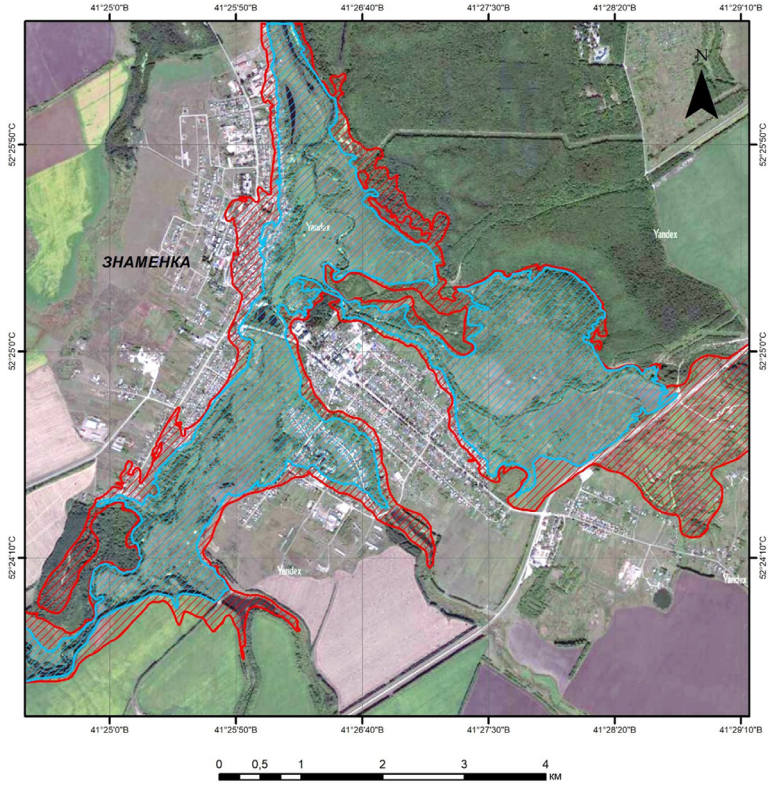


Рис. 7.7. Еще один пример построения буферных зон уже средствами ГИС-редактора: определение зон затопления территории на основе цифровой модели рельефа и дешифрирования данных дистанционного зондирования. Решение подобных задач особенно важно в ходе историко-экологических исследований

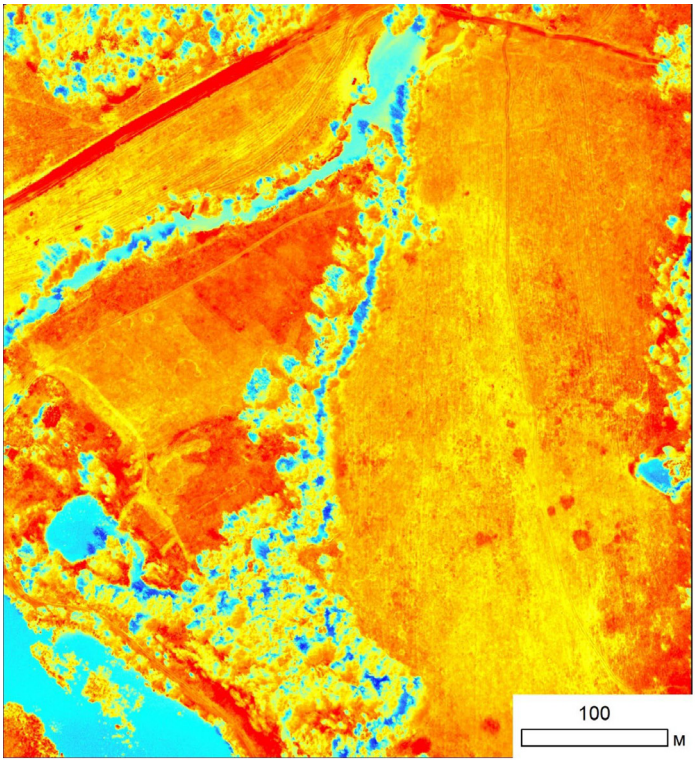


Рис. 7.8. Пример обработки данных дистанционного зондирования Земли путем калькуляции каналов видимого спектра. Комбинация (Red-Green)/(Red+Green). В области высоких значений пребывает почва песчаных типов и верхние сухие слои грунта. Позволяет идентифицировать локальные участки вмешательства в верхние слои или разницу в увлажнении обрабатываемых открытых участков пашни

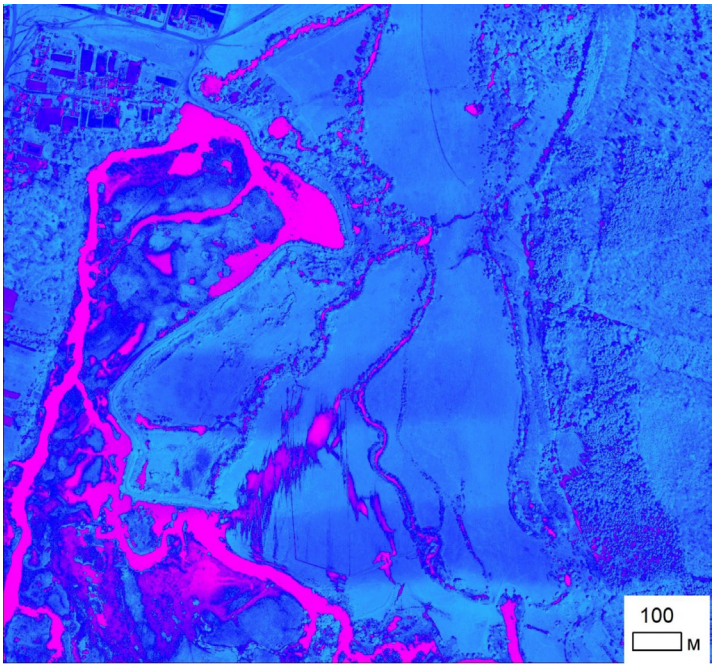


Рис. 7.9. Пример использования ДЗЗ демонстрирует комбинацию (NIR+Red)/NIR. В зону высоких значений попадают открытая вода, снег и тень. Данная комбинация наиболее удобна для картографирования и изучения мелких притоков, ручьев и болотистой местности

Рис. 7.10. На карте путешествия Ш.-Ж. Ромма по Крыму в 1786 г. показан пример построения маршрута по нарративному источнику: после адаптации картографической основы наносятся опорные точки маршрута, которые затем соединяются с указанием направления движения и показателей перемещения

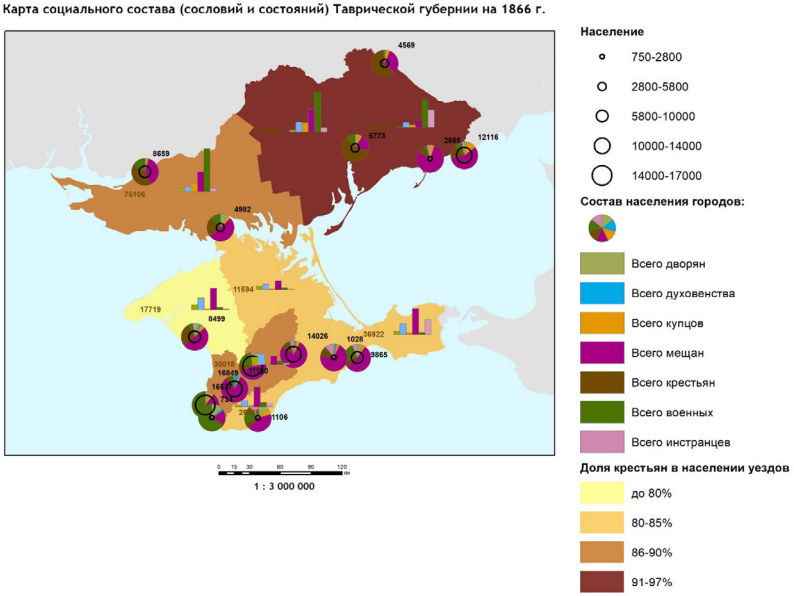


Рис. 7.11. На карте представлена комплексная картограмма, в которой приведены показатели как по площадным объектам, так и по точечным. Применены круговые и столбчатые диаграммы в сочетании с заливкой по площади, что позволяет совместить большой объем информации на одном макете

Глава 8

3D-моделирование, виртуальные реконструкции и VR/AR/MR-технологии в задачах сохранения культурного наследия

(*Л. И. Бородкин*)

Первые приложения технологий 3D-моделирования в гуманитарных исследованиях датируются рубежом 1980–1990-х годов, когда стало возможным разрабатывать 3D-модели и виртуальные реконструкции объектов культурного наследия. Неудивительно, что в мире гуманитарных наук первые шаги в данном направлении сделали в эти годы археологи. Пол Рейли, молодой археолог из Саутгемптонского университета в Англии (рис. 8.1), опубликовал тогда ряд пионерских работ по этой теме, например: «К виртуальной археологии. Компьютерные приложения в археологии», «Трехмерное моделирование и первичные археологические данные» в сборнике «Археология и информационная эпоха». Отметим, что в 2021 году д-р Пол Рейли выступил с докладом на международной конференции по виртуальной археологии, которая проходила в Сибирском федеральном университете в Красноярске.

Распространение технологий трехмерной визуализации в исследовании и сохранении культурного наследия привело в 2009 году к принятию Лондонской хартии компьютерной визуализации культурного наследия. Этот документ установил, в частности, принципы использования цифровой визуализации объектов культурного наследия, методы обеспечения их интеллектуальной целостности,

надежности, адекватного документирования, устойчивости и доступа к виртуальным объектам.

За истекшие 30 лет 3D-технологии шагнули далеко вперед, а визуальный и цифровой повороты, происходившие в исторической науке в начале XXI века, существенно повысили интерес к виртуальным реконструкциям культурного и индустриального наследия. В настоящее время набор этих технологий включает виртуальные трехмерные (3D) интерактивные модели, технологии лазерного сканирования, фотограмметрии, виртуальной (VR), дополненной (AR) и смешанной (MR) реальности, аппаратуру их воспроизведения и интерфейс, специализированное программное обеспечение (в частности, трехмерные редакторы и конверторы 3D-моделей, 3D-движки), компьютерную томографию, LIDAR-технологии и др.

В данной главе мы рассмотрим, какие практики сформировались в области цифрового сохранения трехмерных объектов культурного наследия, уделив особое внимание утраченным (частично или полностью) объектам. Тем самым в центре нашего внимания будут задачи создания *виртуальной реконструкции* таких объектов, которые могут иметь комплексный характер, включая, например, историческую городскую застройку или усадебный комплекс. Важно, что проекты такого рода должны быть полидисциплинарными, ориентированными не только на реконструкцию изучаемого объекта культурного наследия, но и на восстановление его истории, социально-культурного контекста.

Отметим сразу, что можно выделить два основных типа проектов в данной области. Проекты первого типа имеют экскурсионно-познавательные цели; они, как правило, не претендуют на высокую степень достоверности предлагаемой цифровой реконструкции, которая может иметь при этом привлекательную визуализацию, полученную с использованием новейших 3D-технологий. Они дают приблизительное представление о рассматриваемом объекте, без серьезной проработки источниковой базы. В большинстве случаев основные разработчики — это специализированные IT-фирмы, с участием краеведов или сотрудников музеев, а также консультантов по истории архитектуры и медиатехнологий.

Проекты второго типа ориентированы на построение научно обоснованной виртуальной реконструкции, которая базируется на комплексе выявленных разновидовых источников, подвергнутых источниковедческому анализу. Такие реконструкции (назовем их «академическими») могут стать предметом научной публикации

и, конечно, могут использоваться и в экскурсионно-познавательных целях, а также в музейных экспозициях, при реставрации или восстановлении утраченных или руинированных объектов.

Именно проекты второго типа находятся в центре нашего внимания. Материал данной главы содержит достаточно подробные описания ряда таких проектов, что должно способствовать более конкретному пониманию используемых сегодня практик создания виртуальных исторических реконструкций.

Этапы создания виртуальной реконструкции

Процесс создания виртуальной исторической реконструкции объекта культурного наследия условно можно поделить на несколько этапов.

Первый этап заключается в постановке задачи исторической реконструкции, изучении истории объекта реконструкции и определении его значимости.

Второй этап реконструкции — определение круга доступных источников. Это графические, изобразительные, картографические источники и фотографии, а также описательные источники. Нередко важную роль играют источники научно-технической информации (чертежи, планы строений и др.). На основе сформированной источниковой базы будет происходить создание трехмерных моделей. В случае нехватки источников для создания достоверной реконструкции можно обратиться к аналогам, выбор которых должен тщательно обосновываться. В любом случае роль аналогов в сформированной базе источников не должна быть доминирующей. Иначе построенную виртуальную реконструкцию вряд ли можно считать достаточно достоверной, научно обоснованной, но она может быть полезной в иллюстративных и познавательных целях.

Третий этап работы — это критика источников. На этом этапе исследователи, собрав данные, проверяют их на достоверность и находят «нестыковки» в источниках, так как иногда они противоречивы. Например, источники могут отражать изменения в облике реконструируемого объекта. Изменения относятся к различным временным срезам, что требует соответствующей атрибуции, датировки.

Четвертый этап — выбор программного обеспечения для осуществления поставленных задач. Этот вопрос будет рассмотрен ниже.

Пятый этап виртуальной реконструкции заключается в построении трехмерных моделей изучаемого объекта. На этом этапе анализа и обработки информации источников привлекаются не только архитекторы, программисты, но и историки, археологи, краеведы и другие специалисты, которые имеют экспертные знания, полезные в восстановлении облика того или иного артефакта, строения или ландшафта местности.

Шестой этап — текстурирование 3D-моделей, т.е. придание более реалистичного и насыщенного вида рельефным поверхностям моделируемых объектов. Реализация данной функции достигается с помощью специальных библиотек, имеющихся в программах 3D-моделирования.

Седьмой этап — рендеринг (отрисовка), получение качественной визуализации построенной 3D-модели. Современные программы рендеринга могут создавать фотореалистичную визуализацию.

Восьмой этап — создание дополнительных условий для пользователя построенной виртуальной реконструкции. Речь идет о разработке виртуальной 3D-панорамы 360 градусов, виртуального тура, интерактивной системы навигации пользователя в трехмерном пространстве и обеспечении возможности верификации элементов 3D-моделей в ходе такого тура. Интерес пользователей вызывает и ролик, который дает общее впечатление об объекте (комплексе) реконструкции, особенно когда он обладает сложной пространственной структурой. Это может быть, например, дворянская усадьба, которая включает ряд строений и парковую зону.

Новые возможности «погружения» пользователя в воссозданное виртуальное пространство дает применение технологий виртуальной и дополненной реальности (VR/AR).

Важно отметить, что создание трехмерной модели объекта реконструкции может потребовать не только программного обеспечения. В зависимости от конкретной задачи, могут понадобиться и аппаратные технологии, включая лазерные сканеры, модули аэрофотосъемки и фотограмметрию. Поговорим о них подробнее.

Фотограмме́трия — это научно-техническая область, которая ориентирована на разработку методов определения форм, размеров, пространственного положения и степени изменения во времени

различных пространственных объектов по результатам измерений их фотографических изображений.

Характеристики объекта могут изучаться по его изображению на одиночном снимке или по ряду перекрывающихся снимков, которые получены из различных точек пространства. С необходимостью анализа пространства посредством технологий фотограмметрии исследователь сталкивается при работе с фотографиями. Особенно в тех случаях, когда они являются единственным (или главным) историческим источником, который характеризует облик строения целиком. Анализ пространственной перспективы на фотографии, выявление размеров запечатленного строения невозможны без использования технологий фотограмметрии. Такие технологии мы можем найти в специализированном программном обеспечении (например, в пакете PhotoModeler Scanner и его аналогах).

Еще одна активно развивающаяся технология — это лазерное сканирование. Лазерный сканер, или 3D-сканер, это аппаратное устройство, которое анализирует физический объект и на основе полученных данных создает его 3D-модель. Трехмерная модель сканируемого артефакта или строения обычно представляется в виде облака точек или готовой трехмерной модели. Лазерные сканеры появились относительно недавно — в 1990-х годах. Практика современных зарубежных археологических экспедиций в большинстве случаев не обходится без лазерного сканера, несмотря на его дороговизну.

И, наконец, аэрофотосъемка. В задачах виртуальной реконструкции данные аэрофотосъемки позволяют в ряде случаев создать трехмерную модель ландшафта и выступить в качестве отправного материала плана территории, где фотография местности с воздуха позволяет уточнить место расположения объекта в пространстве. В задаче построения виртуальной реконструкции, как правило, большую роль играют плановые аэрофотоснимки со спутников или специальных пилотируемых или беспилотных самолетов, вертолетов, БПЛА. Частный случай использования лазерного сканирования — применение технологий Light Detection and Randing, или LIDAR. Это лазерное сканирование воздушного базирования, которое основано на измерениях расстояния и точной ориентации этих измерений между сенсором и отражающей поверхностью при археологическом и ландшафтном обследовании. Например, технология LIDAR позволяет выявить и четко идентифицировать руины, скрытые под плотными кустарниками или в лесистой местности.

Программное обеспечение 3D-моделирования и создания виртуальных реконструкций

Реконструкция объектов историко-культурного наследия может осуществляться с помощью различных программ 3D-моделирования. Их выбор определяется прежде всего особенностями объекта реконструкции, используемых источников и требований к качеству реконструкции. Рассмотрим кратко характеристику этих инструментов. Мы будем основываться на опыте реализации проектов по созданию 3D-моделей и виртуальных реконструкций на кафедре исторической информатики МГУ.

Отметим, что эти инструменты в основном довольно универсальны. В основном они, конечно, не ориентировались специально на работу с объектами культурного наследия. Изначально они создавались как для профессионального архитектурного моделирования, так и для производства компьютерных игр и 3D-мультипликации. Как правило, создание виртуальной реконструкции требует использования нескольких программ. Одни программы включают решения для непосредственного моделирования объектов — например, 3Ds Max, Maya, Blender, SketchUp, Revit, ArchiCAD. Другие программы позволяют сделать визуализацию с возможностью создания развернутой интерактивной среды — Unity, Unigine, Unreal Engine. Есть также программы с большим уклоном в сторону производства роликов, видеофильмов — Twinmotion, Cinema 4D.

Некоторые программные продукты являются бесплатными, для некоторых существуют специальные предложения для студентов и преподавателей. Таким образом, выбор решения для реконструкции напрямую зависит от предпочтений реконструктора. Программа SketchUp — одна из самых простых и удобных для построения 3D-моделей. Сейчас есть две версии этой программы. Это свободно распространяемая версия продукта SketchUp Make, в которой сохранены все возможности моделирования, но имеются существенные ограничения в функциях импорта-экспорта, и коммерческая версия SketchUp Pro, в которой доступен полный набор функций.

В исторических реконструкциях инструменты SketchUp нашли широкое применение. Для них характерно как создание объектов, которые подлежат реконструкции, так и использование стандартных объектов для заполнения сцен из библиотеки 3D Warehouse. Так, масштабный проект по реконструкции Страстного монастыря,

выполненный на историческом факультете МГУ, был реализован в основном с использованием средств SketchUp. Важно отметить, что SketchUp не был единственной программой для моделирования, а применялся только для создания определенных объектов. Его применение определялось скорее удобством для исполнителей, поэтому для большого проекта частичное использование SketchUp имеет смысл. Таким образом, SketchUp довольно удобен в моделировании несложных объектов. Тем не менее его универсальность в ряде случаев подвергается сомнению, а потому его применение как единственного инструмента для 3D-моделирования сложных объектов — не лучшее решение.

Следующий вариант программы для реконструкций — это Blender. Эта бесплатная программа была разработана на языке Python. Ее интерфейс несложен, расположение панелей инструментов довольно логично. Исторические реконструкции, проводимые с помощью Blender, существуют как на любительском, так и на профессиональном уровне. Можно получить весьма точную модель при наличии достаточного количества фотографий. Также Blender примечателен возможностью внутреннего рендеринга (визуализации). У этой программы есть один недостаток — системные ошибки при работе, даже несмотря на то, что она постоянно совершенствуется.

Следующим решением для реконструкции стали программы для BIM-проектирования (BIM — *Building Information Model*). Две наиболее известные из этих программ — Revit и ArchiCAD. Две конкурирующие компании представляют схожий функционал с отличающимся интерфейсом. Профессиональные версии программного обеспечения довольно дорогие, но в каждом указанном случае существует бесплатная учебная лицензия на один год для некоммерческого использования. Таким образом, использование каждой из этих программ для работы над проектом по реконструкции возможно.

В исторических реконструкциях Revit находит применение в качестве инструмента моделирования как экстерьера, так и интерьера, программу используют и для построения 3D-модели с помощью облака точек. Работа с облаками точек — важная особенность BIM-обеспечения, так как по сканам объекта можно получить наиболее точную и приближенную к реальности 3D-модель.

Программный пакет ArchiCAD[1] имеет интерфейс, который не отличается доступностью и рассчитан, скорее, на профессионального пользователя.

В упомянутом выше проекте по реконструкции Страстного монастыря ArchiCAD нашел активное применение. Так как в проекте участвовал профессиональный архитектор, его совместная работа с историками позволила создать высокоточные модели исторических зданий, а техническая составляющая на момент выполнения проекта способствовала созданию качественной визуализации.

Заключительным пунктом в приведенном списке стала программа 3Ds Max, широко известная в среде 3D-моделлеров. Ее особенность — широкий набор функций, который позволяет эффективно работать с моделями. Возможность создания и обработки как высокополигональных, так и низкополигональных моделей дает исследователю удобный инструментарий для работы с 3D-материалами. Интерфейс 3Ds Max достаточно прост. Для обучения и преподавания предлагается бесплатная лицензия сроком на один год с возможностью продления, что также очень удобно для проведения исследований. Широкие возможности импорта и экспорта этой программы позволяют работать с моделями различных форматов и направлять результаты практически в любую виртуальную среду.

В качестве инструмента визуализации построенных моделей часто выбирают среду разработки Unreal Engine, а также Twinmotion.

Трехмерное моделирование в исследованиях по исторической урбанистике

Цифровые технологии дают новый импульс развитию исторической урбанистики. Сейчас стало возможным создавать виртуальные реконструкции исторической городской застройки в ее эволюции. История городов нередко связана с изменением облика зданий, улиц и площадей. Иногда они исчезают с карты города — в ходе радикальных перестроек или, например, в результате войн или пожаров. Виртуальные реконструкции создаются

[1] ArchiCAD — программный пакет для архитекторов, основанный на BIM-технологии и предназначенный для проектирования архитектурно-строительных конструкций, элементов ландшафта и т.п.

также в сфере сохранения таких объектов культурного наследия, как древние городища, археологические артефакты, дворянские усадьбы и другие.

Определенную специфику в разработку виртуальной реконструкции исторической городской застройки вносит фактор времени — источники отражают эволюцию этой застройки. В этой связи большое значение имеет их датировка.

Один из наиболее известных проектов в этой области — Rome Reborn. Как заявили создатели этой масштабной виртуальной реконструкции, цель их проекта — «демократизировать знания». Проект развивается с 1997 года и представляет собой цифровую реконструкцию Древнего Рима в том виде, который он имел в 320 году нашей эры. Текущая версия модели оптимизирована для использования в режиме реального времени посредством дополненной виртуальной реальности. Пользователь может посетить римский форум с множеством возможностей выбора или поэкспериментировать непосредственно с виртуальной реальностью (рис. 8.2).

С конца 1990-х годов виртуальные реконструкции культурного наследия создаются и в России. Такие проекты реализованы группами исследователей в научных центрах Екатеринбурга, Калининграда, Красноярска, Москвы, Перми, Санкт-Петербурга, Тамбова и др. Пожалуй, наиболее масштабные проекты такого рода реализованы на базе кафедры исторической информатики исторического факультета МГУ. Вот некоторые примеры таких проектов, которые были реализованы в период с 2011 по 2022 год: Страстной монастырь (1654–1937 гг.) и Страстная площадь (XVIII — начало XX в.), исторический ландшафт Белого города (центр Москвы, XVI–XVIII вв.), руинированные подмосковные исторические усадьбы (XVIII — начало XX в.). Отметим, что в течение последних 10 лет по этим проектам на кафедре защищены 16 выпускных работ бакалавров и магистров, подготовлены к защите три работы аспирантов, защищена кандидатская диссертация.

В ряде проектов впервые в отечественной историографии рассматривается опыт использования технологий виртуальной и дополненной реальности в ходе реконструкции утраченных объектов исторической городской застройки, проводится апробация методики и технологии валидации/верификации результатов построенной виртуальной реконструкции, углубления возможностей репрезентации и визуализации этих результатов.

В современных условиях динамичного развития городов, изменяющегося городского ландшафта возрастает интерес к возможностям 3D-технологий, позволяющим создавать виртуальные реконструкции утраченной исторической городской застройки, ее эволюции, нередко изменяющей облик зданий, улиц и площадей в ходе радикальных перестроек или в результате войн, пожаров, стихийных бедствий.

Среди объектов виртуальных реконструкций заметное место занимают монастырские комплексы. В XX веке в России была разрушена значительная часть монастырей, многие из которых представляют интерес для изучения не только с точки зрения архитектурных особенностей, но и с точки зрения их социокультурной роли и экономического значения. В ряде случаев уничтоженные монастыри располагались в Москве. Важным критерием для выбора утраченного монастырского комплекса в качестве объекта виртуальной реконструкции является, наряду с его архитектурными достоинствами, степень сохранности источниковой базы.

Отметим, что число зарубежных проектов по виртуальной реконструкции монастырских комплексов пока невелико, хотя и имеет тенденцию к росту. Это, прежде всего, известные виртуальные реконструкции цистерианского монастыря Санта-Мария XVI в. (район Санзедаш, Португалия), монастыря Санта-Мария XII в. (г. Риполь, Испания), монастыря Сент-Ави Сениер XII в. (департамент Дордонь, Франция), аббатства Клюни X в. (департамент Сона и Луара, Франция), монастыря Христа Пантепопта XI в. (г. Стамбул, Турция). Эти проекты выполнены в XXI веке.

Широкий интерес вызвала работа французских исследователей, построивших виртуальную реконструкцию Собора Парижской Богоматери (Notre-Dame de Paris), восстановив его облик на 13 временных срезах, охватывающих соответствующие этапы расширения и перестройки собора за девять веков его истории. В этой работе, опубликованной в 2013 году, использовались передовые технологии трехмерного моделирования, основанные на лазерном сканировании, что позволило получить плотное облако точек поверхности всего собора (более миллиарда опорных точек).

Что касается российских исследований в этой области, то можно отметить проект «Виртуальная реконструкция Спасо-Преображенского мужского монастыря г. Енисейска XIX в.» Гуманитарного института Сибирского федерального университета [http://www. yeniseisk-heritage.ru] и проекты виртуальной реконструкции

облика монастырей XVII–XIX веков Москвы, выполненные на кафедре исторической информатики МГУ в 2012–2023 годах[1].

Речь идет о московских монастырях, утраченных полностью или частично в 1920–30-х годах. Один из них — Страстной монастырь (московский женский монастырь), основанный в 1654 году и просуществовавший до 1937 года, когда он был полностью разрушен (сейчас на том месте — Пушкинская площадь). В 2014–2016 годах на кафедре при поддержке гранта РНФ был реализован исследовательский проект по виртуальной реконструкции этого монастыря и прилегавшей к нему исторической городской застройки конца XVI — начала XX века. Этот проект имеет выраженный полидисциплинарный характер: творческий коллектив включал историков, искусствоведов, реставраторов, архитектора, IT-специалистов, краеведа и др.

Обратимся к опыту реализации двух проектов виртуальной реконструкции исторической городской застройки Москвы — с тем чтобы рассмотреть предметно основные аспекты масштабных исторических реконструкций. Здесь изложение базируется на публикациях участников этих проектов.

Страстной монастырь. Анализ эволюции пространственной инфраструктуры Страстного монастыря проводился на основе комплекса источников, характеризующих объекты реконструкции на трех временных срезах (рубеж XVII–XVIII вв., 1830-е гг., 1910-е гг. — это позволяет говорить о 4D-моделировании) с учетом социального контекста монастырской жизни и изменявшейся архитектурной среды Страстной площади (см. http://www.hist.msu.ru/Strastnoy/). Пространственная эволюция этой исторической застройки получила отражение в целом комплексе разнообразных источников. К ним относятся проекты архитекторов, планы и чертежи основных

[1] Жеребятьев Д. И. Методы трехмерного компьютерного моделирования в задачах исторической реконструкции монастырских комплексов Москвы. М.: Макс Пресс, 2014. 224 с.; Бородкин Л. И., Жеребятьев Д. И. Технологии 3D-моделирования в изучении пространственных аспектов городской истории: виртуальная реконструкция монастырского комплекса XIX — начала XX вв. // Вестник РФФИ. 2016. № 3 (91). С. 47–60; Мироненко М. С. Современные подходы к 3d-реконструкции объектов культурного наследия: проблемы визуализации и восприятия (на примере Московского Страстного монастыря и Чудова монастыря Московского Кремля) // Электронный научно-образовательный журнал «История». 2015. Т. 6. Вып. 8 (41); Жеребятьев Д. И., Ким О. Г. Особенности виртуальной реконструкции московского Страстного монастыря и прилегающей площади XVII — начала XVIII вв. // Электронный научно-образовательный журнал «История». 2015. Т. 6. Вып. 8 (41).

построек комплекса, делопроизводственные материалы, документы, связанные с перестройкой, реконструкцией, реставрацией и другими изменениями внешнего облика зданий монастыря, гравюры и другие изобразительные материалы, а также фотографии конца XIX — начала XX века.

В ходе создания виртуальной реконструкции решались задачи источниковедческого синтеза, включая выяснение последовательности возникновения источников, их сопоставления по степени достоверности и точности, полноты представления необходимой информации, устранения возможных противоречий источниковой информации. При воссоздании внешнего облика монастырского храма учитывались также архивные описательные документы. Расположение в пространстве каждого объекта реконструкции определялось соотнесением его с планами территории Страстного монастыря 1757, 1773, 1831 годов, а также со сводными топографическими картами, созданными на основе архивных документов участниками проекта (рис. 8.3). На территории площади, прилегавшей к Страстному монастырю, в 1830-х годах существовало более 20 владений, 3D-модели которых в реконструкции привязаны к Генеральному плану Страстной площади 1831 года. Существенно, что совокупность источников по каждому зданию включает чертежи фасадов. Это позволило воссоздать виртуальные модели с документальной точностью.

Одним из ключевых объектов монастыря является колокольня Страстного монастыря. Старое здание колокольни, просуществовавшее вплоть до 1850 года, запечатлено на нескольких графических и изобразительных источниках. Наиболее подробным из них является чертеж колокольни и гравюра лицевого фасада монастыря после пожара 1773 года, выявленные в РГИА. На основе имеющихся источников в программе SketchUp была произведена виртуальная реконструкция старой колокольни Страстного монастыря (она нашла отражение на рис. 8.5).

Отметим, что по отдельным строениям до нас дошли только их планы, фотоизображения конца XIX — начала XX века и гравюры того же времени. Одним из таких строений Страстной площади является храм Дмитрия Солунского. К сожалению, гравюры рубежа XVIII–XIX веков дают достаточно смутное представление об архитектурных формах храма, искажая размеры и пропорции отдельных его элементов. В таком случае наиболее информативным источником выступают старые фотографии здания, которые более точно передают его облик, хотя и относятся к более позднему времени.

Опираясь на данные гравюр и текстовые упоминания о перестройках здания, из имеющихся фотографий делается «вычет» тех составных частей здания и деталей, которые позднее были пристроены. Здесь оптимальным программным обеспечением выступает программа SketchUp. Так, благодаря наличию фотограмметрического инструмента анализа перспективы фотографии и параметров строений MatchPhoto, определив уровень горизонта и указав определенные параметры перспективы в графических источниках, можно рассчитать угол съемки здания фотографом, а затем и размеры всех строений (рис. 8.5). Количество загруженных фотоизображений при наличии разных ракурсов съемки и «реперной точки» непосредственно влияет на точность полученного результата. Реконструкция перспективы для всех графических изображений и определение размера «реперной точки» позволяет задать масштаб реконструкции. Полученная в итоге реконструкция храма Дмитрия Солунского представлена на рис. 8.6.

Преимуществом построенных в нашем проекте 3D-моделей монастырского комплекса является возможность интерактивного просмотра созданной виртуальной реконструкции в онлайн-режиме и ее верификации. Предложенная в проекте процедура верификации подразумевает возможность взаимодействия пользователя с представленными на сайте источниками реконструкции и созданной на их основе виртуальной 3D-моделью, с подробным описанием методики ее построения применительно к каждому зданию. Тем самым повышается источниковедческая достоверность 3D-реконструкции. Здесь возникает и новая источниковедческая задача — презентация всех источников, использовавшихся для восстановления рассматриваемого фрагмента комплекса (с соответствующей критикой источников). По сути, речь идет о создании системы электронной документации виртуальной реконструкции монастырского комплекса. Эта задача решена в нашем проекте на основе разработанного программного модуля.

Построенная компьютерная реконструкция трехвековой эволюции монастырского комплекса и окружавшей его исторической городской застройки Страстной площади показывает те новые возможности в развитии исторической урбанистики, которые открылись перед историками в контексте визуального и пространственного поворотов в сочетании с цифровым поворотом в структуре исторического знания. Полученные результаты представлены в открытом доступе на сайте исторического факультета МГУ (http://hist.msu.

ru/Strastnoy/), что дает возможность пользователю ознакомиться с источниковой базой исследования и построенной виртуальной реконструкцией, представленной с помощью современных средств 3D-визуализации, включая цифровое видео обзора монастырского комплекса и Страстной площади на ранних и поздних временных срезах (см. рис. 8.4, 8.6–8.9), визуальные эффекты дополненной реальности (рис. 8.10) и т.д. В отличие от многих доступных иллюстративных 3D-реконструкций объектов культурного наследия, имеющих в основном экскурсионно-познавательный интерес, данный проект основан на твердой источниковой базе, с использованием верифицируемых методик.

Виртуальная реальность. На втором этапе этой работы (поддержанной фондом «История Отечества») в центре нашего внимания была адаптация технологий виртуальной и дополненной реальности (VR/AR) для создания новых возможностей валидации/верификации результатов построенной виртуальной реконструкции облика Страстного монастыря, углубления возможностей репрезентации и визуализации этих результатов. На данном этапе междисциплинарное исследование проводилось кафедрой исторической информатики МГУ совместно с лабораторией математического обеспечения имитационных динамических систем (МОИДС) механико-математического факультета МГУ.

В качестве дополнительной возможности «погрузиться» в историческое прошлое были созданы исторические панорамы для использования их в планшете или смартфоне. Подразумевается, что пользователь находится на месте утраченного объекта исторической застройки и, наведя планшет на любую точку сегодняшнего городского ландшафта, видит с помощью AR, как выглядело это место 100 или 200 лет назад. Появляется возможность изучения исторической застройки Страстной площади и монастыря с помощью AR и первой версии реконструкции, когда ранее существовавшие элементы исторической застройки «вырастали» на месте сегодняшнего городского ландшафта. Разработанные в МГУ шлем виртуальной реальности и специальные средства отслеживания движений пользователя, работающие совместно с панорамной системой виртуальной реальности, позволяют осуществить виртуальный тур по центру исторической Москвы с «погружением» в историческую городскую среду.

В чем заключается задача использования технологий виртуальной, дополненной и смешанной реальности (VR, AR, MR) в проектах создания виртуальной реконструкции культурного наследия?

Главное — эти технологии позволяют пользователю подключить к восприятию визуального мира прошлого не только зрительный канал, но и остальные каналы восприятия, которые связаны с другими органами чувств. Тем самым можно ощутить иммерсивные эффекты. Такие технологии позволяют накладывать дополнительные слои графики и трехмерные структуры «поверх» объектов окружающего мира, которые пользователь может «рассматривать» через камеру смартфона или специальных очков.

Эти технологии впервые использовались при подготовке космонавтов в еще в середине 1990-х годов. Сегодня VR/AR/MR-симуляторы в ходу у представителей многих профессий.

Как мы «видим» в виртуальной реальности? Главный инструмент здесь — шлемы. Сегодня существует шлемы для смартфонов, для ПК и консолей, есть и автономные шлемы. Созданные приложения на компьютере можно запускать как на привычных смартфонах, так и на специально созданных компьютерах с встроенными дисплеями — в очках и шлемах.

Отдельный интерес представляет адаптация технологий виртуальной и дополненной реальности для создания новых возможностей верификации результатов построенной виртуальной реконструкции и углубления возможностей репрезентации и визуализации этих результатов.

Пользователь «передвигается» в виртуальной воссозданной исторической среде, совершая виртуальный тур. При этом он может приблизиться к интересующему его зданию и, «дотронувшись» до него, получить доступ к соответствующим историческим источникам. Например, это чертежи, планы, старые фотографии, изобразительные материалы, карты, текстовые материалы. Таким образом, он может «верифицировать» достоверность виртуального объекта, наблюдая его в воссозданном пространстве (рис. 8.11).

Так решается новая задача: в ходе перемещений пользователя в виртуальной среде ему обеспечивается доступ не только к самой реконструкции, но и к источниковой базе, на основе которой создается эта реконструкция. Для ознакомления с ней и ее верификации используется виртуальный пространственный интерфейс, который был разработан в МГУ. Такое использование технологий повышает мотивацию при изучении истории культуры и смежных дисциплин. Одна из первых разработок такого рода была впервые осуществлена в рамках исследовательского проекта историков МГУ по виртуальной реконструкции Страстного монастыря.

Реализация технологии виртуальной реальности включает следующие основные шаги. Сначала компьютер создает образ, это трехмерное изображение объектов в виртуальной среде. Затем программная система отображения транслирует его на органы чувств пользователя. Далее установленные на теле пользователя датчики передают компьютеру информацию о действиях и движениях пользователя. Например, о повороте головы, движениях рук или изменении его положения в пространстве. Потом компьютер использует эту информацию для изменения генерируемой им виртуальной реальности и ее образа, который передается на органы чувств пользователя. VR-оборудование, то есть гарнитуры, позволяет пользователю взаимодействовать с виртуальной реальностью, погружаться в нее, совершать передвижения, видеть и слышать виртуальный мир.

Результаты данного этапа проекта, характеризующего возможности нового подхода в исследованиях по исторической урбанистике, представлены на федеральном историко-документальном просветительском портале http://portal.historyrussia.org

Белый город. В полном объеме эта методика была реализована в проекте МГУ по виртуальной реконструкции Белого города, исторической территории, расположенной в центре Москвы XVI–XVIII веков (рис. 8.12). Проект «Пространственная реконструкция исторического ландшафта Белого города Москвы XVI–XVIII вв. (с использованием современных информационных технологий)», поддержанный грантом РФФИ, также имел полидисциплинарный характер, объединяя 11 участников различных научных профилей.

Виртуальный тур в Белом городе XVII века позволяет пользователю «пройтись» по Ивановскому монастырю, заглянуть на территорию городской усадьбы знатного человека. Для достижения реалистичной визуализации математиками МГУ был разработан комплекс на основе шлема виртуальной реальности высокого разрешения.

Система визуализации дает информацию о состоянии математической модели пользователя и окружающих его объектов виртуального мира. Это позволяет моделировать взаимодействие в режиме реального времени. Так достигается полное покрытие поля зрения пользователя и замена реальной среды виртуальной моделью. Для передачи передвижения реального человека в виртуальном пространстве применяется система отслеживания движения. Подобный подход соответствует функции смешанной реальности.

Система смешанной реальности представляет собой программно-аппаратный комплекс, который выполняет функцию визуализации ближайшего окружения пользователя с возможностью отслеживания его движений и передачи этих движений в виртуальное пространство. Эта технология создает иллюзию абсолютного физического присутствия человека в виртуальном пространстве. Таким образом, уже сегодня указанные технологии позволяют использовать иммерсивные возможности для научно верифицируемых исторических реконструкций.

Остановимся подробнее на описании проекта по Белому городу.

Принятое ЮНЕСКО определение исторического городского ландшафта отражает расширенное его понимание, что предполагает многоаспектный и полидисциплинарный подход к его изучению. Здесь возникает исследовательская проблематика, характерная для большинства городов, основанных много столетий назад. Их ландшафт изменялся на протяжении веков под воздействием природных и антропогенных факторов, и его виртуальная реконструкция на тех или иных временных срезах — одна из задач современной исторической урбанистики. Представим одну из первых попыток рассмотрения источниковедческих и методических аспектов виртуальной реконструкции исторического городского ландшафта Белого города.

Целью проекта было объединить данные многочисленных локальных исследований об эволюции градостроительной среды центра Москвы и создать целостную виртуальную реконструкцию исторического городского ландшафта, который изменялся в силу природных и антропогенных факторов. Предполагалось использование современных GIS и 3D-технологий. Хронологические рамки проекта — XVII–XVIII века — определяются возможностями источниковой базы.

В результате проведенной архивной работы были выявлены и оцифрованы источники, характеризующие эволюцию ландшафта Белого города и соответствующей городской застройки рассматриваемого времени. Базовый временной срез для создания виртуальной реконструкции: середина — вторая половина XVIII века, именно этот период обеспечен источниками лучше, чем предшествующий период.

В ходе исследования был проведен источниковедческий анализ выявленных материалов, получена их поливидовая классификация, которая включает изобразительные, картографические, топографические источники, научно-техническую документацию, старые фотографии, а также описательные, текстовые источники, сведения

о владельцах и размерах усадеб на территории Белого города, строительных работах, переписи московских дворов и церковных владений XVII — начала XVIII века, а также переписные книги Москвы и т.д. Немаловажную роль в сформированной базе источников сыграли материалы археологических экспедиций, работавших на территории Белого города, и геолого-геоморфологические данные.

На основе сформированной источниковой базы были решены следующие задачи: создание базы данных по материалам о сооружениях изучаемой территории Белого города; создание виртуальной реконструкции рельефа изучаемой территории Белого города; виртуальная реконструкция доминантных объектов исторической застройки Белого города (Ивановский монастырь, храмы, палаты, строения городских усадеб); размещение (координатная привязка) реконструированных объектов на воссозданном рельефе Белого города; формирование виртуальной среды Белого города для использования VR-технологий в целях «погружения» пользователя в воссозданный исторический городской ландшафт.

Предлагаемая впервые виртуальная реконструкция исторического ландшафта Белого города позволяет оценить роль антропогенного фактора, выявить влияние расширяющейся городской застройки на эволюцию его рельефа и доминатных построек. Решение поставленных задач потребовало использования целого ряда цифровых технологий.

В ходе работы по виртуальной реконструкции исторических сооружений Белого города в 2019 году был использован один из методов получения натурных данных о частично сохранившихся архитектурных объектах — технологии 3D-сканирования и аэрофотосъемки. Для этого применялись современные технические средства — лазерный сканер, коптер. В качестве примера 3D-моделей, выполненных разными методами, в данной работе представлены реконструкции ряда объектов исторической застройки восточной части Белого города. Это церковь князя Владимира в Старых Садех XVII–XVIII веков, палаты князей Голицыных — княгини Щербатовой, думного дьяка Украинцева (рис. 8.13–8.16). Выявленные источники позволяют предложить верифицируемые реконструкции их внешнего облика. В рамках проекта было проведено лазерное 3D-сканирование фасадов храма Владимира в Старых Садех, существующего в наше время. Проводилась также комбинированная съемка храма с последующим построением облака из 73 млн точек по материалам аэрофотосъемки, что позволило построить 3D-модель

этого храма (рис. 8.17). Виртуальная реконструкция храма XVII века была создана на основе его сложившегося облика, путем воссоздания его ретроспективной эволюции (рис. 8.18). В качестве основного программного средства для построения виртуальной реконструкции строений Белого города использовалась программа ArchiCAD.

Следует отметить, что рельеф местности восточной части Белого города довольно сложный, перепад высот здесь достигает нескольких десятков метров, что потребовало создания трехмерной карты этого рельефа с помощью программных возможностей пакета QGIS. Особенности рельефа тщательно учитывались в ходе построения виртуальной реконструкции Ивановского монастыря. Это основной доминантный объект на территории Белого города, расположенного на Ивановской горке. Особое внимание уделялось храму монастыря. Построенная визуализация 3D-моделей строений монастыря и всего комплекса в целом воспроизводит облик монастыря XVIII века (рис. 8.19), который радикально изменился после Отечественной войны 1812 года. Спустя полвека по проекту архитектора М. Д. Быковского был построен новый монастырский комплекс, сохранившийся до настоящего времени.

Воссозданная виртуальная среда исторического ландшафта Белого города дала возможность для апробации предложенных в ходе проекта алгоритмов отслеживания движений человека, синхронизации реальных движений человека и их визуальной имитации в условиях виртуальной реальности исторического города. Был разработан программный модуль верификации созданных 3D-моделей. Пользователь при этом получает прямой доступ к соответствующим историческим источникам, он передвигается в условиях виртуальной реальности, использует VR-технологии, чтобы «погрузиться» в воссозданный исторический городской ландшафт.

Основные результаты, полученные в ходе проекта, опубликованы[1] и представлены (включая виртуальный тур, маршрут которого задает пользователь) на сайте проекта http://www.landscape.vrmsu.ru/

[1] Бородкин Л. И. О виртуальной реконструкции исторического городского ландшафта Белого города // Историческая информатика. 2019. № 4. С. 90–96; Лемак С. С., Чертополохов В. А., Кручинина А. П., Белоусова М. Д., Бородкин Л. И., Мироненко М. С. Задачи оптимизации расположения элементов интерфейса в виртуальной реальности (в контексте создания виртуальной реконструкции исторического рельефа Белого города) // Историческая информатика. 2020. № 1. С. 81–93; Ким О. Г., Моор В. В., Жеребятьев Д. И. Виртуальная реконструкция доминантных объектов исторической застройки Белого города Москвы (XVI–XVIII вв.) // Историческая информатика. 2020. № 2. С. 100–134; Чернов С. З.,

Виртуальная реконструкция индустриального наследия

Постиндустриальное развитие приводит к завершению жизненного цикла большого количества индустриальных объектов, предприятий, которые оказываются невостребованными ввиду изменения производственных технологий, возросших экологических требований и, таким образом, утери этими объектами их первоначальных функций. Дискутируемым остается вопрос о том, как должна развиваться дальнейшая судьба бывших промышленных зданий и их оборудования. Одна из новаций в этой области связана с применением технологий 3D-реконструкций и виртуальной реальности в изучении и сохранении памятников промышленной эпохи, особенно в случае невозможности их физического сохранения. В России первопроходцами здесь стали уральские исследователи Е. А. Курлаев и Ю. М. Баранов, которые создали в начале 2000-х годов виртуальную 3D-реконструкцию уральского металлургического завода XVIII века.

Рассмотрим проблемы виртуального сохранения объектов индустриального наследия на примере создания научно обоснованной 3D-реконструкции объектов Трехгорного пивоваренного завода, который берет свое начало в 1875 году. Завод, крупнейшее предприятие этой отрасли не только в Москве, но и в России, в настоящее время заброшен и постепенно разрушается. Рассмотрим проблемы цифрового сохранения индустриального наследия на примере исследования А. А. Гасанова, проводимого в рамках проектов кафедры исторической информатики МГУ[1].

Основным объектом являлся производственный корпус «Варня» Трехгорного пивоваренного завода, его внешний облик, существенно изменившийся в XX веке, и его оборудование, а также производственные процессы. Варня, то есть варочное отделение завода,

---

Бойцов И. А. Археологические источники визуальной реконструкции исторического ландшафта восточной части Белого города Москвы (XIV–XVI вв.). Ивановская горка. // Историческая информатика. 2020. № 2. С. 135–178.

[1] Гасанов А. А. Виртуальная реконструкция индустриального наследия: опыт 3D-реконструкции архитектурного облика производственного корпуса Трехгорного пивоваренного завода в Москве рубежа XIX–XX вв. // Историческая информатика. 2021. № 2. С. 88–114; Гасанов А. А. Создание интерактивных сред и использование технологий виртуальной реальности в реконструкции производственных процессов (на примере Трехгорного пивоваренного завода в Москве на рубеже XIX–XX вв.) // Историческая информатика. 2021. № 3. С. 69–85.

является необходимой частью любого пивоваренного предприятия и центральным звеном пивоваренного процесса.

Создание 3D-модели Варни включало моделирование геометрии здания, наложение текстур, материалов и финальную визуализацию (рис. 8.20). Для создания геометрии модели был выбран 3Ds-max. В качестве ключевой технологии выбрана Camera Match, позволяющая совместить двумерное изображение объекта с 3D-пространством, что облегчает расчеты расположения и масштаба элементов объекта и делает 3D-реконструкцию более точной. Для текстурирования и последующей визуализации был выбран игровой движок Unreal Engine 4. Использовался также метод тайлингового наложения текстур, предполагающий наложение текстур на объект, с многократным повторением одной текстуры без видимых швов.

Виртуальная реконструкция внутренних помещений и оборудования корпуса Варни на рубеже XIX–XX веков проводилась в основном по фотоматериалам и чертежам. Визуализация потребовала работу с движком Unreal Engine 4 и настройку освещения. Использование VR-технологий для моделирования производственного процесса позволяет пользователю стать виртуальным участником этого процесса, познать особенности одной из промышленных профессий прошлого.

Для финальной презентации виртуальной реконструкции в формате интерактивного VR-приложения были созданы системы информационных справок и верификации исторических источников. Обе они заключались в выводе в зону видимости пользователя поясняющей информации, в первом случае — сообщений о необходимых на данном этапе действиях и назначении тех или иных объектов, во втором — изображений источников, на которых основан данный элемент реконструкции. В рамках виртуального тура пользователь имеет возможность ознакомиться с источниковой базой реконструкции, провести ее визуальную верификацию.

Иллюстрации «погружения» пользователя в виртуальную среду технологического процесса конца XIX века представлены на рис. 8.21–8.25.

В ходе рассмотренного проекта была получена интерактивная система в виртуальной реальности. Она позволяет пользователю проводить в виртуальной среде операции технологического процесса на предприятии XIX века, сделать это в интерьере Варни Трехгорного завода. Подобный формат виртуальной реконструкции может найти

в будущем более широкое применение в реконструкции технологических процессов прошлого.

* * *

Разумеется, базовые знания о возможностях создания 3D-моделей и виртуальных реконструкций культурного наследия не исчерпываются изложенным материалом. Полезным будет ознакомление с опытом реализации проектов, реализованных в этой развивающейся междисциплинарной области, представленным в публикациях журнала «Историческая информатика» последнего десятилетия.

Иллюстрации подготовлены при участии А. А. Гасанова, Д. И. Жеребятьева, В. В. Моора, М. С. Мироненко

Рис. 8.1. Пол Рейли

Рис. 8.2. Одна из реконструкций из проекта Rome Reborn

Рис. 8.3. Модель расположения зданий Страстного монастыря на плане местности. Начало XVIII в.

Рис. 8.4. Визуализация виртуальной реконструкции Страстного монастыря и окружающей городской застройки на рубеже XVII–XVIII вв.

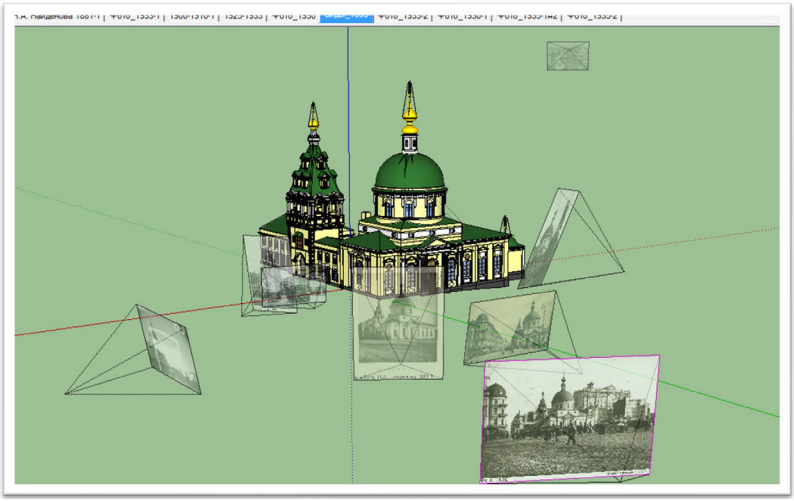


Рис. 8.5. Расчет точек фотосъемки храма Дмитрия Солунского (технология фотограмметрии)

Рис. 8.6. Визуализация виртуальной реконструкции храма Дмитрия Солунского и Страстной площади. 1830 г.

Рис. 8.7. Визуализация виртуальной реконструкции Страстного монастыря и окружающей городской застройки с высоты птичьего полета. 1910 г.

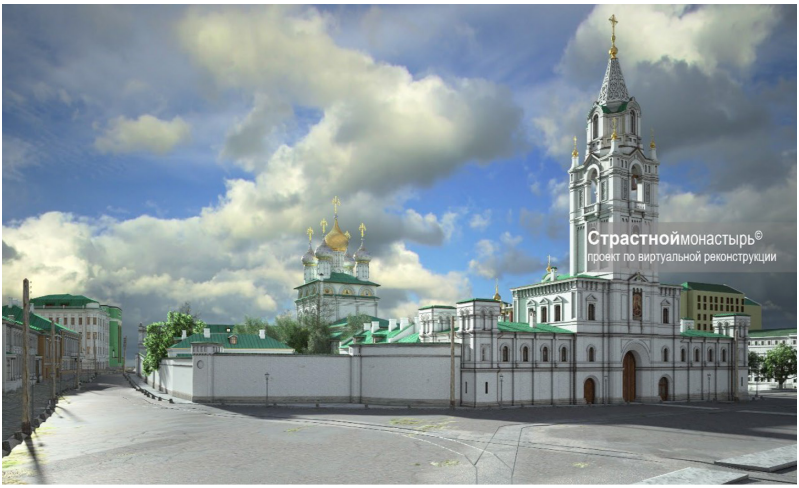


Рис. 8.8. Визуализация виртуальной реконструкции комплекса Страстного монастыря. 1910 г.

Рис. 8.9. Визуализация виртуальной реконструкции колокольни Страстного монастыря и окружающей городской застройки. 1910 г.

Рис. 8.10. Дополненная реальность: схематичное отображение в AR виртуальной реконструкции Страстного монастыря на существующую городскую застройку Пушкинской площади

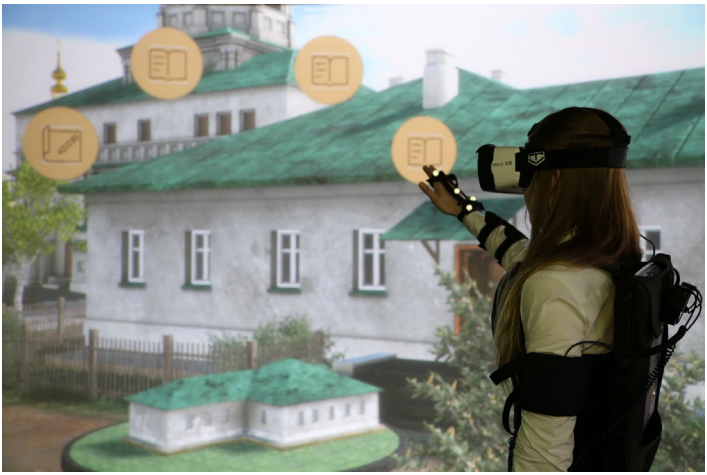


Рис. 8.11. Визуализация интерфейса VR-модуля верификации на панорамной системе виртуальной реальности (в кружках показаны логотипы видов источников)

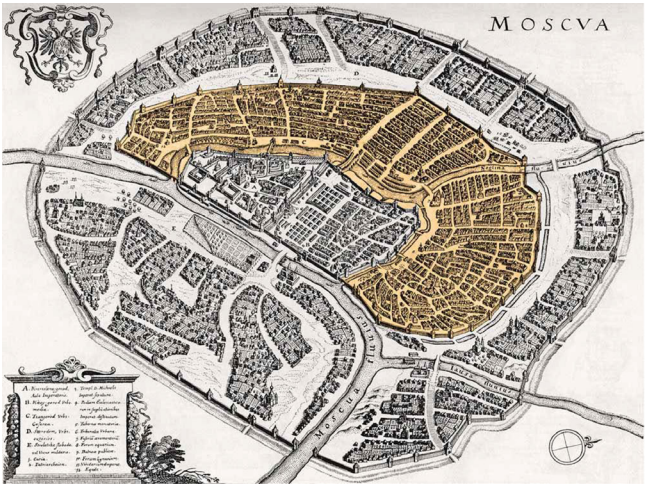


Рис. 8.12. Белый город в XVII в. (территория выделена цветом)

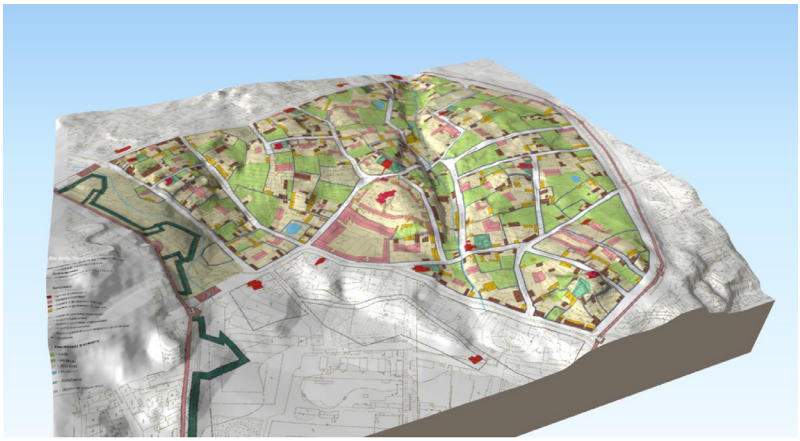


Рис. 8.13. Виртуальная 3D-реконструкция ландшафта восточной части Белого города, отражающая его парцелляцию на вторую половину XVIII в. (построена в программе QGIS)

Рис. 8.19. Визуализация 3D-моделей строений Ивановского монастыря. XVIII в.

Рис. 8.17. Лазерное сканирование храма Святого Владимира в Старых Садех в программе Cloud Compare (пример недоступных участков сканирования)

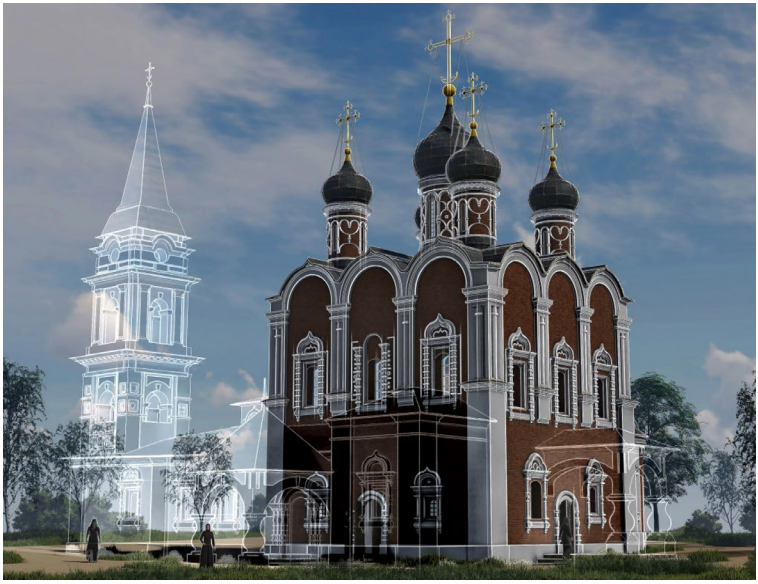


Рис. 8.18. Церковь князя Владимира в Старых Садех. Вид с юго-запада. Виртуальная реконструкция на XVII в.

Рис. 8.14. Палаты кн. Н.С. Щербатовой с домовой церковью Казанской Божьей Матери. Виртуальная реконструкция на конец 1750-х гг.

Рис. 8.15. Визуализация виртуальной реконструкции палат Голицыных. 1769 г.

Рис. 8.16. Палаты думского дьяка Украинцева и его хозяйственные постройки. Середина XVIII в.

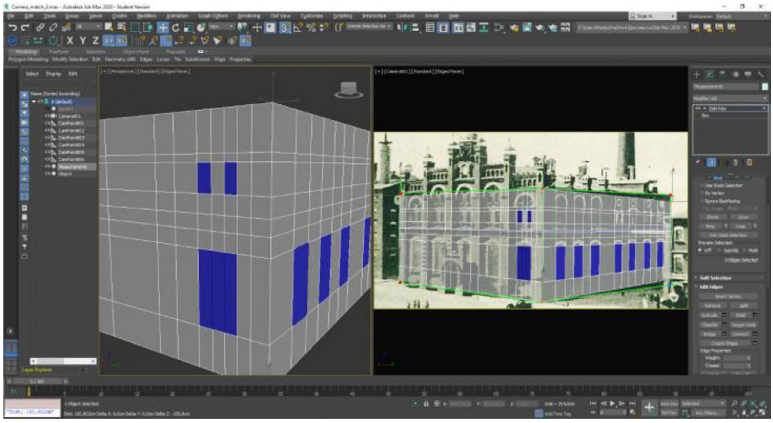


Рис. 8.20. Виртуальная реконструкция Варни — производственного здания Трехгорного завода. Применение технологии Camera Match в 3Ds-max. Синим выделены очертания окон, границы архитектурных элементов обозначены линиями

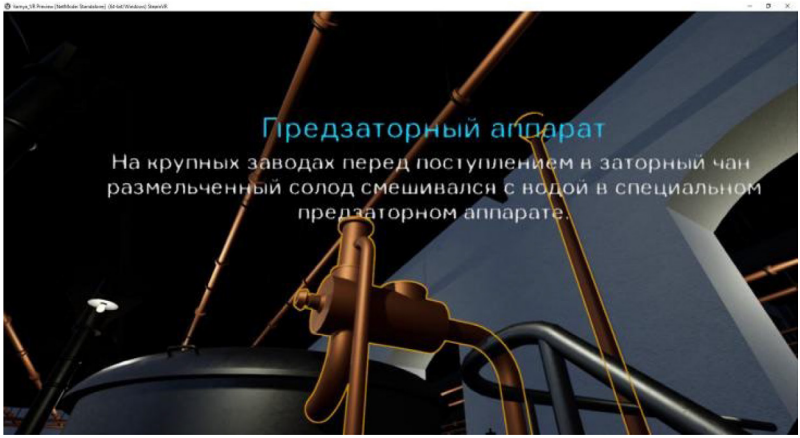


Рис. 8.21. Пользователь получает подсказку о предзаторном аппарате. При нажатии на кнопку контроллера текстовое описание сменится на изображение источника

Рис. 8.22. Пользователь видит изображение предзаторного аппарата на источнике

Рис. 8.23. Технологический процесс требует кипячения солода. Пользователю предлагается растопить печь дровами. При этом объекты, с которыми требуется взаимодействие, подсвечены цветной обводкой. По кнопке контроллера обводку и подсказки можно отключить

Рис. 8.24. Операция выщелачивания дробины. Пользователь имеет возможность разобраться в назначении и принципе работы этого устройства

Рис. 8.25. Рендеры внутренних помещений дают высококачественную визуализацию внутренних помещений Варни, с показом технологической линии и основного оборудования

Глава 9

Сетевой анализ данных (social network analysis, SNA): подходы и технологии

(*А. В. Сметанин*)

Почему социально-сетевой анализ называется социальным?

В самом начале предлагаю обратиться к определению «социальной сети». Авторы новейшего издания определяют социальную сеть как «набор акторов или других сущностей (entitles), а также набор или наборы отношений, привязанных к ним»[1]. Подобное формальное определение воспроизводится в большинстве справочников. Попытаемся привести примеры, подходящие под данное определение. Класс школьников — идеальный пример, дети являются акторами, выражают симпатии и антипатии, готовность к соучастию, а дружба или помощь в учебе выступают в роли связей. Один из самых ранних, еще интуитивных выходов на анализ социальных сетей связан именно со школьниками. Немецкий учитель начальных классов Йоханнес Делич в течение 1880–1881 учебного года фиксировал формирование дружеских отношений между своими учениками и пытался понять, почему какие-то дети заняли центральное положение в сети[2]. К слову, этими детьми оказались второгодники, отличники и мальчик, угощавший всех конфетами. Сам Делич

[1] Knoke D., Yang S. Social network analysis. Third edition. Thousand Oaks, SAGE. 2020. P. 1.

[2] Delitsch J. Über Schülerfreundschaften in einer Volksschule // Zeitschrift für Kinderforschung. 1900. № 5. P. 150–162.

понятие «сеть» не использовал и даже не пытался нарисовать ее, ему для анализа хватило таблицы.

Другой пример — сеть стран, участвующих в финансировании различных гуманитарных проектов в Африке. Можно ли такую сеть назвать социальной? Как правило, исследователи отвечают, что можно. Международные отношения существуют только в социальных системах, а страны являются своеобразными акторами. Третий пример — сеть книг на сайте онлайн-магазина, составленная на основе интересов пользователей. Чем чаще посетители покупали две книги одновременно, тем выше шанс, что эти книги будут связаны. И здесь мы встречаемся с проблемой. Книги, составляющие основу сети, акторами с точки зрения социологии не являются точно, и неясно, насколько правомерно называть это социальной сетью. В целом, среди сторонников социально-сетевого анализа нет единства по вопросу о том, что является допустимым предметом исследования, поэтому иногда уместнее говорить просто о сетевом анализе, либо скорректировать определение социальной сети. Лучшее определение дал Стив Боргатти. Сеть — это «способ мышления о социальных системах»[1], т.е. сеть — это то, что исследователь воспринимает как сеть.

Сеть как концепт не является монопольным достоянием социально-сетевого анализа. Исследование сетей характерно для структуралистских направлений социологии, теории семантических сетей в лингвистике, моделирования транспортных систем и т.д. Как справедливо замечает Л. И. Бородкин[2], элементы сетевого анализа можно найти и в работах советских историков, например, уже в 1970-е годы моделировались сети сходства (генеалогические схемы) различных списков летописей.

Закончить социологическое введение необходимо важной оговоркой. Один из пионеров применения социально-сетевого анализа в отечественной науке Г. В. Градосельская отмечает, что сетевой анализ оперирует одновременно понятием актора (т.е. действующего лица) и понятием структуры, которая объективно ограничивает возможности социального действия[3]. Как следствие, обращаясь к се-

[1] Borgatti S. P., Everett M. G., Johnson J. C. Analyzing Social Networks. Sage Publications, 2013. P. 17.

[2] Бородкин Л. И. Сетевой анализ в исторических исследованиях: микро-макроподходы // Историческая информатика. 2017. № 1. С. 110–124.

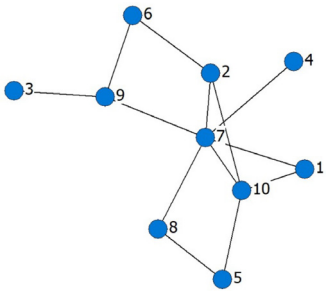
[3] Градосельская Г. В. Сетевые измерения в социологии: учеб. пособие / под ред. Г. С. Батыгина. М., 2004. С. 8–9.

тевому анализу, исследователь априори соглашается с тем, что сеть реальна, а позиция актора в сети задает рамки для его активности.

Анатомия сетей

Существует несколько обстоятельных очерков развития метода[1], к которым стоит адресовать интересующихся. Эта история не была линейна, в ней были важные акторы, но не было отца-основателя. Теоретическим фундаментом явился формальный подход Г. Зиммеля и выросшие из него направления структурной социологии. Как правило, историю собственно метода отсчитывают с 1930-х годов и социограмм Я. Морено[2]. Длительное время визуальный анализ сетей оставался ключевой аналитической техникой. Обратимся и мы к данной технике для усвоения базовых понятий сетевого анализа.

Визуальное представление сети называется *графом*. Теория графов начала формирование более чем на столетие раньше социально-сетевого анализа, и до 1960-х годов эти направления развивались параллельно. Как следствие, одни и те же структурные элементы сети в социологии и геометрии могут называться по-разному. Рассмотрим случайно сгенерированный граф.



О том, что в социально-сетевом анализе может выступать в роли акторов, шла речь выше. Стоит добавить, что в специальной

[1] Freeman L.C. The Development of Social Network Analysis. A Study in the Sociology of Science. Vancouver, 2004; Scott J. Social Network Analysis. Fourth edition. SAGE Publications, 2017. P. 11–41.

[2] Moreno J. Who Shall Survive? A new Approach to the Problem of Human Interrelations. Washington, 1934.

литературе синонимами «актора» выступают математические понятия *узел* и *вершина* (node, actor, vertix). Аналогами понятия «связь» выступают термины *ребро* и *дуга* (tie, arc, edge, link).

Выделение связей является наиболее трудоемкой и творческой частью сбора данных. Связи могут фиксировать как относительно статичные явления, например уважение, схожесть политических взглядов, цитирование, родство, так и явления динамические. К последним относятся транзакции, акты помощи, общение, распространение информации и т.д. Достаточно интересным направлением является построение сетевых моделей на основе корреляционных матриц. В таком случае сетевой анализ является продвинутым вариантом метода корреляционных плеяд[1]. Всегда существует риск, что полных данных о связях собрать не удастся, что негативно повлияет на качество модели и все расчеты.

К структурным элементам сети относятся *диады* — пары связанных узлов, число диад равно числу связей в сети, в нашем примере их 13. *Триады* могут быть полными, например, треугольник, составленный из узлов 2–7–10. Также триады могут быть неполными (при отсутствии одной связи), например, это структура из узлов 3–9–6. К протяженным структурам сети относятся *пути*, т.е. расстояния от одного актора до другого. Они измеряются количеством «пройденных ребер». Между акторами 3 и 4 существует два пути — длинный, равный пяти ребрам (3–9–6–2–7–4), а также короткий, равный трем ребрам (3–9–7–4). Кратчайший путь называется *геодезическим*. Также популярно исследование эго-сетей, т.е. совокупности узлов, напрямую связанных с одним избранным узлом. Например, *эго-сеть* узла 9 составлена из узлов 3, 6, 2, 7 и связей между ними.

Вернемся к визуальному анализу сети. Какие выводы можно сделать, не прибегая к расчетам? Однозначно, фиксируется центральная роль актора 7, также к ядру сети можно отнести весь треугольник 2–7–10. Если пытаться выделить группы, то наиболее логичным выглядит разделение на верхнюю и нижнюю половину. В целом же связи выглядят разреженными. Хотя последний вывод не столь однозначен — он будет справедлив для сети друзей, но вряд ли мы назовем связи разреженными, будь это сеть долгов между коллегами.

Потенциал исключительно визуального анализа ограничен, и лучшим доказательством является история развития метода.

[1] Бондарева Е. В., Стеценко Н. В. Метод корреляционных плеяд в практике педагогических исследований // Математическая физика и компьютерное моделирование. Т. 21. № 2. С. 52–58.

По словам Л. Фримана, 1940–1960-е годы стали «темными веками» социально-сетевого анализа[1], когда после яркого старта в 1930-е годы наблюдалось охлаждение к возможностям метода. Основная причина заключалась в том, что большинство работ заканчивалось отрисовкой графа и его экспертным описанием, не было ясности, а что именно дает представление социальной реальности в виде сети. В начале 1970-х годов математика стала недостающим языком описания сетей.

Построение и типология сетей

Социально-сетевой анализ оперирует всего двумя структурными элементами (узлы и связи между ними), поэтому задача построения графа сводится к тому, какими образом закодировать связи или их отсутствие между каждой парой акторов. В теории графов существует два основных способа решения данной задачи.

Матрица инцидентности — это способ построения графа, где задаются отношения между акторами (строки) и ребрами, т.е. связями (столбцы). Рассмотрим матрицу творческих связей голливудских актеров 1940–1950-х годов. Основанием для выделения связи является совместное участие хотя бы в одном полнометражном фильме. Наличие связи, как и в матричных операциях, кодируется единицей. Обозначения E1 — E7 в данном случае являются названиями семи обнаруженных связей (от *edge*, т.е. ребро).

	E1	E2	E3	E4	E5	E6	E7
Вивьен Ли	1				1		
Марлон Брандо	1					1	
Грейс Келли			1	1			
Хамфри Богарт		1					
Одри Хепберн		1					1
Фрэнк Синатра			1			1	1
Кларк Гейбл				1	1		

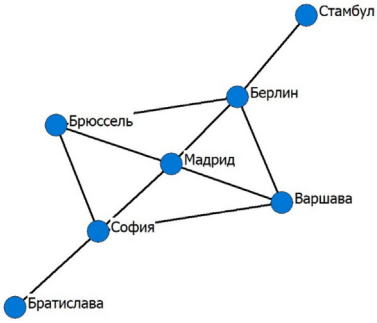
[1] Freeman L.C. The Development of Social Network Analysis. A Study in the Sociology of Science. Vancouver, 2004. P. 64.

Несложно заметить, что в каждом столбце только две единицы, потому что одно ребро может связывать лишь двух акторов. Так, ребро E3 появилось благодаря съемкам Грейс Келли и Фрэнка Синатры в фильме «Высшее общество». Данная матрица тождественна приведенному графу. Для большей наглядности названия ребер подписаны.

В практике гуманитарных исследований гораздо более удобным и простым способом моделирования являются *матрицы смежности*. Более того, именно такие матрицы по умолчанию используются в программном обеспечении. Основной принцип построения таблицы — это фиксация связи между двумя акторами, один из которых записан в строке, а другой в столбце. Принципиально важно, чтобы названия строк и столбцов были зеркальным отражением друг друга: названия должны идти ровно в том же порядке и в абсолютно идентичном написании, иначе программы распознают одного актора как двух разных. Ниже представлена таблица побратимства некоторых городов Европы.

	Мадрид	Варшава	Стамбул	Берлин	Брюссель	София	Братислава
Мадрид							
Варшава	1						
Стамбул							
Берлин	1	1	1				
Брюссель	1			1			
София	1	1			1		
Братислава						1	

Поскольку в таблице всегда существует две ячейки на пересечении нужных акторов, то обычно единицу ставят в обе ячейки, либо заполняют только часть таблицы под центральной диагональю (как это сделано в примере). В матрицах смежности допустимы связи актора с самим собой, если это важно для исследования, например, в сетях цитирований так можно обозначить самоцитирование ученого. Граф, соответствующий таблице, расположил в центре Мадрид, что можно было предсказать по общему количеству связей у этого города.

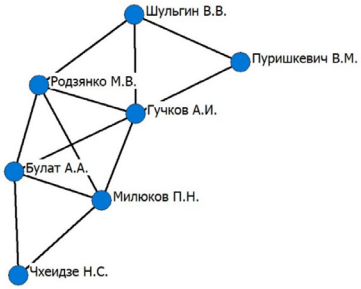


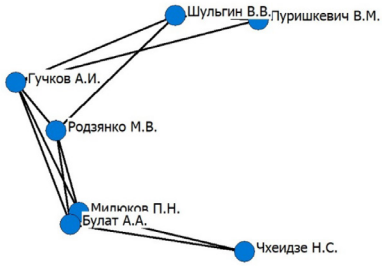
Очевидно, многих исследователей не удовлетворит глубина данных, которую позволяют закодировать представленные выше матрицы. Бинарный способ кодировки (0 или 1) является признаком *невзвешенных графов*, и такой способ достаточно груб для решения многих научных задач. Если исследователь располагает данными о силе связей, то имеет смысл строить *взвешенный граф*. Именно исследование взвешенных графов привело М. Грановеттера к известной теории «силы слабых связей», когда оказалось, что найти престижную работу проще через шапочных знакомых, а не лучших друзей[1]. Моделируется подобная сеть идентичным образом, однако вместо единиц указываются веса связей, т.е. значения, отражающие их силу. Воспользуемся в качестве примера данными о депутатах Государственной Думы Российской империи[2]. Значения в таблице отражают количество законопроектов, подписанных одновременно обоими депутатами в 1907–1912 годах.

[1] Granovetter M. The Strength of Weak Ties // American Journal of Sociology. Vol. 78. Issue 6. 1973.

[2] Сметанин А. В. Институт фракции в Государственной Думе Российской империи: 1906–1917 гг.: дис. … канд. ист. наук. Пермь, 2016.

	Гучков	Милю-ков	Пуриш-кевич	Чхеидзе	Шуль-гин	Родзян-ко	Булат
Гучков							
Милюков	7						
Пуриш-кевич	3	0					
Чхеидзе	0	4	0				
Шульгин	4	0	3	0			
Родзянко	11	3	0	0	2		
Булат	4	14	0	5	0	2	





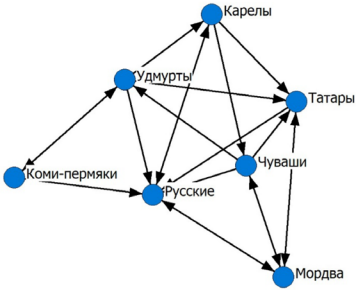
На этот раз рассмотрим сразу два графа, соответствующих таблице. Задача построения графа в социально-сетевом анализе называется *укладкой*. Основная проблема заключается в том, что построить абсолютно корректный граф с нужными длинами ребер в двухмерном пространстве обычно невозможно, и чем больше размер сети, тем больше погрешность. Укладка всегда сопряжена с усреднением и подгонкой связей, поэтому любой программный пакет предлагает несколько способов укладки. Второй (правый) вариант использует метод главных компонент и точнее воспроизводит длины ребер: депутатов с родственными политическими позициями располагает рядом. При этом очевидным образом теряется наглядность.

Сети допускают еще одно усложнение — связям можно придать направление. Такие графы со «стрелками» называются *ориентированными* и кодируются несимметричными матрицами, где в строках задаются узлы, от которых исходит связь, а в столбцах — узлы, на которые эта связь направлена. На этот раз смоделируем сеть

статей Википедии о народах России. Основанием для выделения связи определим наличие в основном тексте статьи об одном этносе упоминание другого этноса. Для простоты смоделируем невзвешенную (бинарную) сеть.

	Русские	Татары	Мордва	Чуваши	Удмурты	Коми-пермяки	Карелы
Русские		0	1	0	0	0	1
Татары	1		1	0	0	0	0
Мордва	1	1		1	0	0	0
Чуваши	1	1	1		1	0	0
Удмурты	1	1	0	0		1	1
Коми-пермяки	1	0	0	0	1		0
Карелы	1	1	0	1	0	0	

На графе заметно, что существуют как *реципрокные* связи, т.е. взаимные, двунаправленные, так и однонаправленные. Например, в статье о карелах упоминаются татары, но не наоборот.

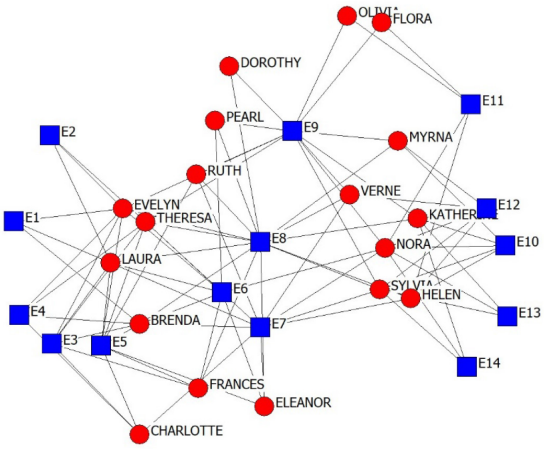


Важно понимать, что эта сеть не тождественна сети генетических связей этносов. Это репрезентация истории и современности народов России на свободно редактируемом электронном ресурсе.

В завершении разговора о построении сетей разберем еще один вопрос. Возможно ли построить сеть, где будет более одного набора акторов? Например, не только литераторы, но и журналы, где они публиковались. Для решения таких задач применяются *двумодальные сети* (*two-mode)* которые в социологии нередко называют *аффилиативными*. В качестве примера стоит ознакомиться с классическим

исследованием 1941 года антрополога Эллисона Дэвиса и его коллег[1]. Сеть, известная под названием «Женщины Юга», в SNA является чем-то вроде сети «по умолчанию» и используется для проверки возможностей новых метрик. Для каждой из включенных в наблюдение 18 женщин фиксировалось посещение 14 значимых для их социального круга мероприятий (Events 1–14). Ниже воспроизводится лишь фрагмент этой сети. Двумодальные сети включают два набора акторов, один заводится через строки таблицы, другой через столбцы. В такой сети актор всегда будет связан только с акторами из другого набора.

	E1	E2	E3	E4	E5	E6	E7
Evelyn	1	1	1	1	1	1	0
Laura	1	1	1	0	1	1	1
Theresa	0	1	1	1	1	1	1
Brenda	1	0	1	1	1	1	1
Charlott	0	0	1	1	1	0	1
Frances	0	0	1	0	1	1	0
Eleanor	0	0	0	0	1	1	1



[1] Davis A., Gardner D., Gardner M.R. Deep South: A Social Anthropological Study of Caste and Class. Chicago and London, 2022. P. 108.

С помощью данной модели авторы доказали, что внутри, казалось бы, гомогенного по составу общества белых женщин маленького городка на Юге США выделяются неформальные клики, а в каждой клике есть люди, составляющие ядро, периферию и просто примыкающие. На полном графе это заметно и без каких бы то ни было расчетов. Кстати, сам Дэвис с коллегами для обработки таблицы использовали не сетевые методы.

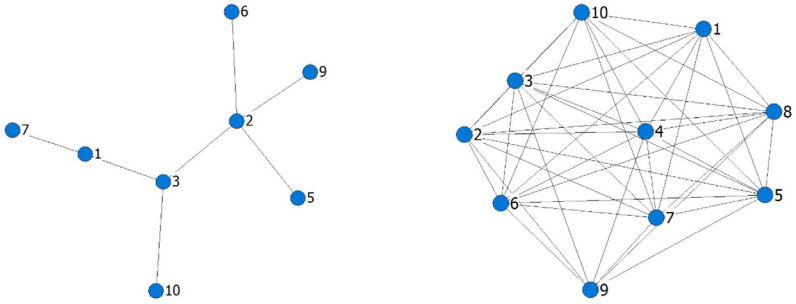
Данный тип сетей в науке получил ограниченное применение в силу того, что двумодальные сети сложнее интерпретировать и метрик для их анализа существует на порядок меньше, чем для одномодальных.

Математика сетей

Все математические метрики в социально-сетевом анализе можно разделить на три группы по сходству решаемых задач. Общесетевые метрики направлены на изучение свойств сети, например сплоченности акторов или степени централизации графа. Метрики центральности определяют структурное положение конкретного актора и важность этого актора в сети по какому-либо параметру. Отдельной задачей является выделение в сети относительно сплоченных групп.

Прежде чем перейти к обзору сетевых метрик, попытаемся ответить на вопрос: будут ли представленные ниже сети интересны для социально-сетевых расчетов?

Безусловно, можно встать на формальную позицию и сказать, что любая сеть может быть предметом для анализа. Особенно



если речь идет о сравнительном исследовании. Например, представьте, что перед вами сети общения рабочих в двух соседних цехах. В таком случае в одном цехе мы наблюдаем минимальные социальные коммуникации, а в другом, наоборот, коллектив друзей. Однако два показателя сигнализируют о том, что результаты математического анализа этих двух конкретных сетей будут не слишком интересны.

Во-первых, *размер сети*, т.е. количество акторов. Нижнего предела для сетевого анализа не существует, и десять единиц вполне допустимое число. Например, при исследовании эго-сетей творческих связей архитекторов Московского метро производился сравнительный анализ компактных графов[1]. Однако интерпретация маленьких сетей не требует обращения к математике, расчеты покажут ровно то, что очевидно и при визуальном анализе. Другая особенность маленьких сетей — изменение или неверное построение одной связи может серьезно переиграть все метрики.

Во-вторых, структура обеих моделей не представляет сложности для интерпретации. Чрезвычайно разреженные сети, равно как и сети «все связаны со всеми», несут скудную информацию о структуре. В первом случае все метрики покажут минимальные значения, во втором, наоборот, близкие к максимуму.

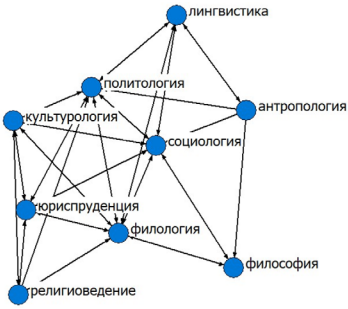
Общесетевые метрики

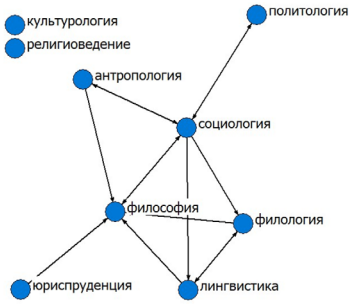
В формате обзорной главы едва ли возможно осветить все метрики и способы их расчета. Более важным представляется понять другое — для оценки одного и того же свойства сети могут использоваться принципиально различные методы. В случае общесетевых метрик чаще всего определяют степень *сплоченности сети* (cohesion), далее разберем две метрики, ведущие к этой цели.

Для проверки используем лингвосемантические сети. Сетевой анализ нередко используется для изучения сетей, составленных из слов какого-либо текстового корпуса. Далее попытаемся установить, в какой степени в русскоязычных текстах связаны различные социальные науки. Нижеследующие графы сформированы на основе

[1] Ермошин А. Д. Сетевой анализ просопографической базы данных об архитекторах Московского метрополитена 1935–1991 гг. // Историческая информатика. 2017. № 4. С. 130–142.

запросов к Национальному корпусу русского языка[1]. Первая сеть составлена на основании смыслового сходства слов, рассчитанного методами дистрибутивной семантики. Связь фиксировалась в том случае, если слово входит в первую десятку семантических ассоциатов другого слова. Вторая сеть основана на принципе коллокации, т.е. частоты взаимной встречаемости слов в одном предложении. Также фиксировались лишь попадания в первую десятку. Поскольку принцип составления матриц был подробно изложен ранее, то ограничимся только представлением графов.





Основной вывод анализа очевиден: сеть смыслового сходства намного более сплоченная, чем сеть взаимной встречаемости. К тому же в сети коллокации слова «культурология» и «религиоведение» не имеют ни одной связи. Теперь выразим эту мысль на языке математики.

Наиболее популярной из метрик сплоченности является показатель *плотности* (density). Она рассчитывается как отношение имеющихся в сети связей к потенциально возможному и принимает значения от 0 (нет ни одной связи) до 1,0 (все узлы связаны со всеми). Иногда плотность выражают в процентах, умножая индексы на 100. Так, для нашей «ассоциативной» сети плотность будет равна 0,56, а для сети взаимной встречаемости слов всего 0,19. Иными словами, в первой сети присутствует 56% всех возможных связей, это очень высокий показатель. В практике работы с большими сетями плотность редко превышает 10%.

[1] Национальный корпус русского языка. 2003–2023 [Электронный ресурс]. URL: ruscorpora.ru (дата обращения: 02.06.2023).

Принципиально иной подход к определению сплоченности предлагает метрика *среднее геодезическое расстояние* (average geodesic distance), которая рассчитывается как средняя длина кратчайших путей между всеми парами узлов в сети. Чем ближе акторы находятся друг к другу, тем меньше будет данный показатель. Расстояние между каждой парой акторов равно количеству пройденных ребер, т.е. минимальный показатель может быть равен 1,0 (между непосредственно связанными акторами находится одно ребро). Среднее геодезическое расстояние в «ассоциативной» сети оказалось равным 1,54, в сети коллокации 1,83. Логично, что чем меньше связей в сети, тем более длинными получаются маршруты.

Согласно одной метрике сплоченность «ассоциативной» сети оказалась выше почти в 3 раза, согласно другой — в 1,2 раза. Какую же метрику стоит выбрать? Безусловно, каждый исследователь подбирает метрики, исходя из поставленных целей анализа. Среднее геодезическое расстояние в большей мере оценивает скорость прохождения информации по сети, что не слишком подходит для оценки близости наук, выраженной в текстах.

Среди других метрик сплоченности стоит обратить внимание на *связность* (connectedness), оценивающую долю пар узлов, между которыми в сети существует путь. Помимо сплоченности, существуют метрики для оценки *транзитивности* (transitivity), т.е. наличия альтернативных вариантов прохождения путей, разнообразие возможных посредников между акторами. Метрики *кластеризации* оценивают однородность сети, наличие в ней участков с разной плотностью. Метрики *централизации* пытаются определить ядро или ядра внутри сети. Овладение любыми метриками требует обращения к учебной литературе.

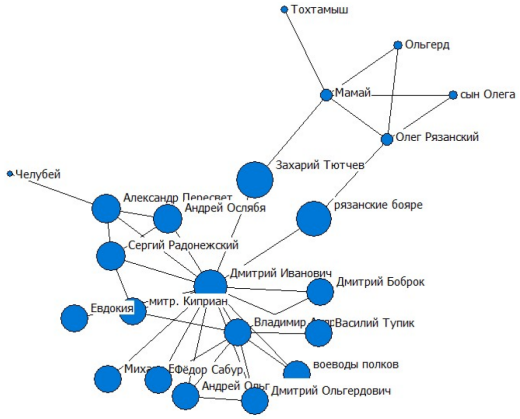
Метрики центральности

Изучение центральности акторов — наиболее проработанная часть социально-сетевого анализа. Обратной стороной такого внимания является неопределенность, с которой сталкиваются исследователи при выборе из десятков метрик. На опытном примере можно убедиться, что само понятие «центральности» приобретает совершенно разный смысл в зависимости от метода ее расчета.

Обратимся к популярному направлению сетевого анализа — моделированию сетей персонажей литературных произведений. Наиболее примечательные проекты в данном направлении реализованы Ф. Фишером и Д. А. Скоринкиным[1], в первую очередь, это Russian Drama Corpus, предлагающий сетевые модели драматических произведений на русском языке[2]. Далее мы обратимся к сети ключевых персонажей литературного памятника XV в. «Сказание о Мамаевом побоище». Основанием для выделения связи является какое-либо непосредственное взаимодействие персонажей (диалог, письмо, бой и т.д.). Сеть невзвешенная, учитывается только факт взаимодействия, а не его частота контактов или сила.

Достаточно часто в качестве метрики центральности используют показатель *степени* узла, т.е. количества связей у актора. На взгляд автора данного очерка, степень является свойством узла, но мало говорит о структурной позиции актора, поэтому использовать степень в качестве метрики центральности стоит с осторожностью.

Без сомнений, к метрикам центральности относится, например, *шаговая доступность*, которую чаще употребляют в англоязычном варианте *k-step–центральность*. Она оценивает количество узлов, до которых возможно добраться через k шагов от исследуемого актора, обычно берут k = 2 («друг моего друга»). Чем больше ты можешь охватить и мобилизовать людей, тем более ты значим для сети — такова

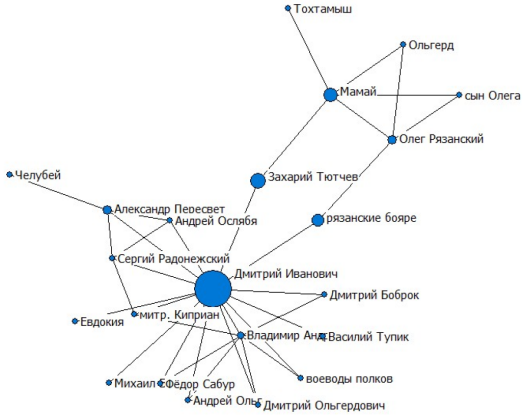


[1] Fischer F., Skorinkin D. Social Network Analysis in Russian Literary Studies // The Palgrave Handbook of Digital Russia Studies. Palgrave Macmillan, 2021. P. 517–536.

[2] Russian Drama Corpus [Электронный ресурс]. URL: https://dracor.org/rus (дата обращения: 02.06.2023).

идея метрики. На представленном графе все тесное окружение Дмитрия Донского получило высокие индексы, тогда как изолированному стану врагов (Мамай, Ольгерд и др.) в центральности было отказано, у них в двухшаговой доступности слишком мало «друзей».

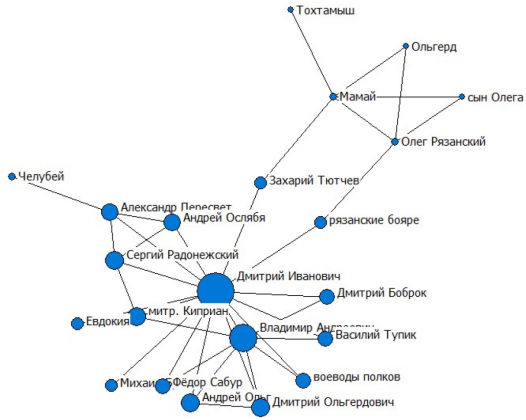
Одной из наиболее популярных метрик сетевого анализа является *центральность по посредничеству* (betweenness centrality). Метрика фиксирует, как часто узел находится на кратчайшем пути между любыми парами узлов в сети, т.е. как часто играет роль оптимального посредника. Высокие индексы получают акторы, располагающиеся в «тонких» местах сети, а также акторы, у которых много *подвесок* (pendants), т.е. узлов, связанных только с одним актором.



Обратите внимание: результаты анализа полностью изменились. Главным «брокером» сети стал Дмитрий Донской, на котором замыкается большинство диалогов в тексте. В своей части графа заметную роль стал играть Мамай, поскольку «обойти» его совсем периферийным персонажам сложно. Данная сеть также позволяет увидеть главную слабость данной метрики — переоценка второстепенных пограничных акторов. Захарий Тютчев был посланником князя Дмитрия к Мамаю и играет эпизодическую роль в повествовании, но структурно для изучаемой метрики он оказался очень важен, что с точки зрения содержания текста несправедливо.

Для сравнения стоит взять еще одну метрику — *центральность по собственному вектору* (eigenvector centrality). Принцип ее вычисления довольно сложен, ограничимся лишь информацией, что рассчитывается она на основании суммы центральностей узлов,

окружающих данный актор[1]. Важность актора определяется по силе его окружения («короля играет свита»).



Из трех метрик только эта особо выделила роль Владимира Андреевича Серпуховского, который в «Сказании» является вторым по важности персонажем, ближайшим сподвижником Дмитрия. Это отнюдь не означает, что центральность по собственному вектору лучше ранее рассмотренных метрик. Известно, что она сильно завышает центральность первого узла в рейтинге, переоценивает центральность плотных участков сети. Стоит лишь повторить уже озвученную мысль: метрики подбираются под исследовательскую задачу, а не перебираются в поисках нужного результата.

Центральные персонажи «Сказания о Мамаевом побоище» в соответствии с разными метриками совпали в минимальной степени.

2step-центральность	Центральность по посредничеству	Центральность по собственному вектору
Захарий Тютчев — 20 Рязанские бояре — 19 Дмитрий Донской — 18 Сергий Радонежский — 16 Пересвет — 16 Ослябя — 16	Дмитрий Донской — 163 Захарий Тютчев — 45 Мамай — 38,5 Рязанские бояре — 30 Пересвет — 20 Олег Рязанский — 19,5	Дмитрий Донской — 0,58 Владимир Андреевич — 0,40 Сергий Радонежский — 0,24 митр. Киприан — 0,24 Дмитрий Ольгердович — 0,24 Андрей Ольгердович — 0,24

[1] Encyclopedia of Social Network Analysis and Mining. Second Edition /ed. R. Alhajj, J. Rokne. Springer, 2018. P. 194.

Исследователь оказывает исключительное влияние на результаты не только выбором метрик. Представьте, что Захарий и рязанские бояре были бы удалены из сети на стадии подготовки таблицы, как второстепенные персонажи, тогда результаты анализа изменились бы кардинально.

Среди других метрик центральности популярна, например, *центральность по близости* (closeness centrality) — сумма кратчайших расстояний от узла до всех остальных узлов сети. Также в научных работах часто применяется *бета-центральность* (Beta centrality) — модифицированный вариант центральности по собственному вектору.

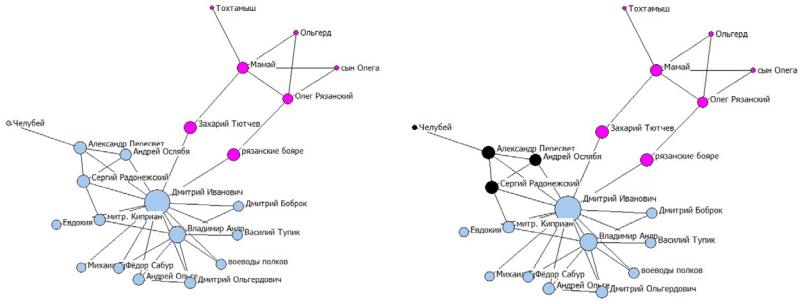
Выделение подгрупп в сети

Если говорить точно, то выделение *подгрупп* (subgroups) относится уже не к метрикам сети, а к алгоритмам. Также сразу стоит внести критическую ноту — алгоритм, который одинаково хорошо работал бы с сетями разных конфигураций, не существует и, по всей видимости, создать его невозможно. Исследователи постоянно сталкиваются с тем, что алгоритмы предлагают абсурдные группировки акторов с точки зрения экспертного знания. Этому есть несколько причин. Во-первых, любой алгоритм основан на каких-либо свойствах сети, акторов или связей, и эта базовая идея может не соответствовать сути авторской модели. Во-вторых, связи в сети могут символизировать несходные явления, и одинаково точно разделить сеть распространения информации и сеть дружеских отношений вряд ли возможно. В-третьих, сами связи в сети могут быть выстроены исследователем достаточно грубо, а для ряда алгоритмов разница между слабой и сильной связью несущественна.

Остановимся подробно на *алгоритме Гирван—Ньюмана*, одном из наиболее известных методов разделения сети на подгруппы. Идея разработчиков была достаточно проста — в сети необходимо отыскать «тонкие» места, где сеть проще всего разорвать, вероятно, именно в таких местах и проходят границы между группами. Для этого вычисляются ребра с самыми высокими индексами посредничества (ровно тем же способом, что и центральность по посредничеству для акторов). Если у какой-то связи индекс посредничества очень высок, значит, связь является *мостом* (bridge), она несет важнейшую

роль для единства сети. Именно такие связи алгоритм и разрушает до тех пор, пока сеть не начнет делиться на группы.

Удобно продемонстрировать действие алгоритма на уже знакомой сети персонажей «Сказания о Мамаевом побоище», поскольку нам точно известно, что в этой сети есть, как минимум, два противоборствующих лагеря. Наиболее качественными вариантами алгоритм полагает разделение на 2, 3, 5 и 8 групп, причем наиболее высоко оценивается разделение на три группы. Качество разбиения традиционно измеряется с помощью процедуры модулярность (modularity). Ниже приведено разбиение сети на 2 и 3 группы.



Наличие группы врагов Дмитрия Донского алгоритм определил достаточно легко, поскольку в сети даже визуально определяются два важнейших моста. Впрочем, Захария Тютчева математика ошибочно отнесла в стан Мамая. Достаточно интересным и логичным получилось решение с тремя группами. Третий обособившийся коллектив связан с сюжетом о Троицком монастыре и его священнослужителях, сюда же попал Челубей, убитый монахом Пересветом на Куликовом поле.

Среди других алгоритмов необходимо упомянуть выделение в сети *клик* (cliques), т.е. групп акторов, где все связаны со всеми. В нашей сети самая большая клика включает четырех князей-командующих (Дмитрий, Владимир и братья Ольгердовичи). К популярным методам также относится разбиение графа на *фракции* (factions), когда сеть делится на заранее заказанное число подгрупп, а далее эти группы переформировываются и уточняются до тех пор, пока не будет найдено наиболее качественное решение.

Необходимо отметить, что исследователи ищут оригинальные способы выделения групп, не привязанные к сложным алгоритмам. Например, в исследовании филологов Б. В. Орехова, П. В. Успенского

и В. В. Файнберг эмигрантский круг общения поэта В. Ходасевича сегментировался с помощью метрик центральности, поиска «пограничных» фигур[1]. Это позволило доказать, что «младшее» поколение эмиграции успешно интегрировалось в литературное сообщество Русского зарубежья.

Существует и обратная операция, когда группы выделяются экспертно, а затем исследуются структурные позиции этих групп. В образцовом исследовании научного сообщества (членов Ассоциации «История и компьютер») И. М. Гарскова с помощью подобного алгоритма выявила особенности формирования региональных и межрегиональных групп[2]. В частности, была определена ключевая роль лидеров региональных сообществ в связывании сети.

Атрибуты и проверка гипотез

В ведущей научной периодике, посвященной социально-сетевому анализу, практически все статьи не останавливаются на фиксации центральности или выделении групп, а обращаются к проверке гипотез. Гипотеза может звучать следующим образом: как на эффективность команды в киберспорте влияет интенсивность социальных взаимодействий и степень централизации команды. Для решения такой задачи необходимо собрать сведения о нескольких командах, в частности, об успешности действий в игре, количестве и силе связей в командной сети, также необходимо сверить эти данные со структурой сети. К слову, это реальное исследование испанских авторов по материалам сообщества League of Legends[3]. Исследователи пришли к выводу, что более успешны команды с интенсивным взаимодействием и не подверженные сильной централизации. Впрочем,

[1] Орехов Б. В., Успенский П. Ф., Файнберг В. В. Цифровые подходы к Камер-фурьерскому журналу В. Ф. Ходасевича // Русская литература. 2018. № 3. С. 19–53.

[2] Гарскова И. М. Сетевой анализ историографии: динамика формирования региональных центров исторической информатики // Историческая информатика. 2017. № 3. С. 94–115; Гарскова И. М. Сетевой анализ историографии: динамика формирования межрегиональной компоненты сети АИК // Историческая информатика. 2017. № 4. С. 112–129.

[3] Mora-Cantallops M., Sicilia M.-A, Team efficiency and network structure: The case of professional League of Legends // Social Networks. Vol. 58. July 2019. P. 105–115.

аналогичное исследование, посвященное крикету, наоборот, показало эффективность команд, замкнутых на одном лидере[1].

Так или иначе, проверка гипотез требует дополнительных данных, которых в самой сети нет. Такие внедренные данные называются *атрибутами*. Например, изучая складывание неформальных сообществ в преподавательской среде нам могут быть важны сведения о возрасте и гендере акторов, профессиональных интересах, принадлежности к научным школам и т.д. Атрибуты невозможно завести в матрицу, где кодируются связи, поэтому в программах социально-сетевого анализа атрибуты заводятся отдельными таблицами с сохранением точного написания имен акторов.

Следуя логике Д. Боргатти и его коллег, гипотезы можно разделить на четыре вида[2]. Для большей наглядности предлагается таблица с примерами.

Виды гипотез	Смысл	Пример
Монадические (monadic)	определение наличия связи между двумя параметрами узлов	связана ли центральность семьи в брачных сетях небольшого городка с богатством этой семьи
Диадические (dyadic)	определение закономерностей в связях между парами узлов (т.е. в диадах)	есть ли зависимость между желанием семей породниться и наличием деловых связей между семьями
Смешанные (dyadic–monadic)	определение связи между параметрами узлов и диадических связей этих же узлов	будут ли семьи примерно равного достатка чаще устраивать браки друг с другом, нежели с более бедными семьями
Общесетевые (network level)	определение наличия связи между двумя параметрами сети	влияет ли средний достаток жителей на плотность брачных связей в маленьких городках

Работа с гипотезами требует от исследователя различного уровня подготовки, где-то достаточно корреляционного анализа двух рядов данных, но в некоторых случаях (особенно это касается смешанных

[1] Mukherjee S. Leadership network and team performance in interactive contests // Social Networks. Vol. 47. Oct. 2016. P. 85–92.

[2] Borgatti S.P., Everett M.G., Johnson, J.C. Analyzing Social Networks. Sage Publications, 2013. P. 125–148.

гипотез) требуется дополнительная работа по преобразованию данных.

Программное обеспечение

Широкому распространению социально-сетевого анализа, помимо прочего, способствует благоприятная ситуация с программным обеспечением. Все ведущие пакеты для анализа являются бесплатными к использованию. Остановимся на наиболее популярных программах среди исследователей.

Pajek — пакет, разработанный в 1997 году словенским ученым Андрем Мрваром. Несмотря на постоянные обновления, программа обладает достаточно архаичным интерфейсом и в настоящее время популярна, в основном, среди исследователей, имеющих дело с большими сетями.

UCInet — разработка специалистов Гарварда и конкретно С. Боргатти. Несмотря на солидный возраст (первая версия — 1992 г.) программа постоянно обновляется. UCInet оснащен обширным инструментарием для обработки сетевых данных, по количеству же аналитических техник данный пакет, вероятно, является лидером. Другое достоинство — наличие учебных материалов, разработанных для данного пакета. Недостатки связаны с устаревшим и часто неудобным интерфейсом. Программа плохо подходит для работы с большими сетями, разработчики не советуют анализировать сети с более чем 5000 акторов. UCInet работает только с системой Windows.

Gephi — программа нового поколения, ведет начало с 2008 года, работает на всех популярных операционных системах. Предлагает более наглядный способ работы, где мощная визуализация является ключевым звеном интерфейса. При этом разнообразие научного инструментария относительно невелико, и, в отличие от UCInet, исследователь не может контролировать многие стадии анализа.

R — язык программирования для статистических расчетов, близкий к Python. Среди подключаемых пакетов можно найти инструменты для сетевого анализа, наиболее популярен пакет igraph. Также для языка R разработан пакет SIENA, самый популярный инструмент для анализа *динамических* (т.е. изменяющихся во времени) сетей. R позволяет коммуницировать сетевой анализ с другими

статистическими методами. Впрочем, для использования указанных возможностей требуется владеть этим языком программирования.

Некоторое неудобство доставляет разнообразие форматов сетевых данных, фактически для каждой программы в 1990–2000-е создавался или выбирался собственный формат. Золотой стандарт не выработан до сих пор.

Какое бы программное обеспечение исследователь ни выбрал, качество анализа зависит, в первую очередь, от правильных решений самого исследователя. Тема должна подразумевать целесообразность использования сети как способа описания реальности. Набор акторов и связей должен быть полон. И, наконец, выбор используемых метрик должен соответствовать целям исследования, а не готовым ответам в голове исследователя.

Глава 10

Информационная инфраструктура цифровых гуманитарных исследований

(*А. Б. Антопольский, А. Ю. Володин*)

Методологические вопросы инфраструктуры DH

Важной организационной и структурной особенностью современной эпохи развития цифровых научных коммуникаций и цифровой науки в целом стала эволюция институций, обеспечивающих функционирование научных организаций и ученых в цифровой среде: библиотек, архивов, издательств, репозиториев, фондов алгоритмов и программ, информационно-коммуникационных сервисов и др. В настоящее время все эти институции получили общее название инфраструктуры цифровой науки. Инфраструктура цифровой науки стала выполнять функции системы научной информации, которая сформировалась в 1950–1960-х годах XX века. В полной мере этот процесс относится и к цифровой гуманитаристике.

Наиболее полно эти структуры развиты в Евросоюзе, где к инфраструктуре цифровой науки принято относить следующие ресурсы и услуги:

- основное научное оборудование, или наборы инструментов;
- коллекции, архивы, или научные данные;
- вычислительные системы и коммуникационные сети;
- любую другую инфраструктуру, открытую для внешних пользователей.

Если говорить об общем направлении, в котором развивается цифровая гуманитаристика во всем мире, то это, конечно, открытая наука. Если в первые годы XXI века концепция открытой науки разрабатывалась в основном странами промышленного Севера, то с 2021 года, после принятия на 41-й сессии генеральной ассамблеи

ЮНЕСКО Рекомендаций ЮНЕСКО по открытой науке[1], концепция открытой науки стала общемировым трендом.

При создании информационных ресурсов, которые являются как источником, так и результатом цифровых гуманитарных исследований, концепция открытой науки предлагает придерживаться принципов FAIR (Findability, Accessibility, Interoperability, Reusability — находимость, доступность, совместимость, повторное использование). Реализация этих принципов требует прежде всего организации разных форм сотрудничества и коллабораций при создании и поддержке информационных ресурсов DH, в том числе между различными институциями, дисциплинами, платформами и технологиями.

Важнейшие значение, особенно для повторного и многократного использования, приобретает устойчивость ресурсов DH. При этом под устойчивостью следует понимать не только и столько технологические аспекты, хотя они также имеют важное значение, сколько организационные, финансовые, кадровые вопросы поддержки ресурсов и систем и, конечно, содержание информации, прежде всего ее полноту, достоверность и актуальность.

Проблема устойчивости критична для всего мира, но для России особенно, поскольку российские инфраструктурные проекты для научной коммуникации практически не институционализированы, если не учитывать традиционные библиотеки и архивы, которые недостаточно адаптированы к условиям цифровой науки.

Можно назвать множество проектов, особенно научных информационных систем, которые были разработаны при грантовой поддержке и прекратили свое существование или развитие после прекращения гранта. Нужно учитывать, что инфраструктурные сервисы почти не могут быть самоокупаемы, а краудфандинг в российских условиях ненадежен. Поэтому основным способом поддержки информационной инфраструктуры в российских условиях должно быть государственное задание. Однако такой подход требует серьезной коррекции системы финансирования в области научной информации.

Одним из компоненентов инфраструктуры DH должна быть система учета информационных ресурсов DH. Такая система начала создаваться в виде справочно-информационной системы

[1] UNESCO Recommendation on Open Science. URL: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

по цифровой гуманитаристике (СИС ЦГ), соглашение о создании которой заключили ИНИОН РАН и Сибирский федеральный университет[1]. Одна из перспективных задача СИС ЦГ — учет правовых возможностей повторного использования ресурсов, инструментов и сервисов, созданных в сфере DH.

Однако перспектива устойчивого развития справочных систем вызывает у специалистов ряд вопросов. Эта проблема подробно рассмотрена в статье[2], а также на семинаре «Цифровая среда» Сибирского федерального университета[3], где обсуждался опыт создания и поддержки каталога инструментов цифровой гуманитаристики DIRT[4]. В частности, речь шла о необходимости институционализации службы каталога или другой подобной справочной системы, о преимуществах и недостатках волонтерского участия в проекте, сочетании единоначалия и коллективного руководства проекта и других организационных аспектах разработки и ведения справочных ресурсов.

Решающее значение для устойчивости инфраструктурных проектов для цифровой гуманитаристики имеет их востребованность. Например, широко известный консорциум TEI[5], который координирует и разрабатывает методы электронного представления текстов, успешно существует и развивается почти 30 лет, его руководства пользуются заслуженной популярностью. На базе методов TEI реализовано свыше 200 проектов.

В качестве положительного примера решения проблемы устойчивости ресурса можно привести так называемый Мэддисоновский проект[6] — база данных, которую собрал Джэймс Мэддисон, и после

---

[1] Антопольский А. Б., Володин А. Ю. Справочно-информационная система по цифровой гуманитаристике: опыт описания интернет-ресурсов российских архивов // Историческая информатика. 2022. № 2. С. 50–66. DOI: 10.7256/2585–7797.2022.2.38236 EDN: HOVPGY URL: https://nbpublish.com/library_read_article.php?id=38236

[2] Grant K., Dombrowski Q., Ranaweera K., Rodriguez-Arenas O., Sinclair S., Rockwell G. Absorbing DiRT: Tool Directories in the Digital Age // Digital Studies/le Champ Numérique. 2020. № 10(1). DOI: http://doi.org/10.16995/dscn.325

[3] Домбровски К. Directories as Utopian Infrastructure (Каталоги как утопическая инфраструктура) // Материалы семинара «Цифровая среда». 16.11.2022 Сибирский федеральный университет. URL: https://dhri.timepad.ru/event/2224905/

[4] В настоящее время каталог DIRT интегрирован с каталогом TAPOR.3.0. URL: https://tapor.ca/pages/about_tapor (доступно 01.09.2023).

[5] The Text Encoding Initiative (TEI) https://tei-c.org/

[6] Groningen Growth and Development Centre Faculty of Economics and Business Maddison Historical Statistics. https://www.rug.nl/ggdc/historicaldevelopment/maddison/ (доступно 20.03.2023).

его смерти данные исправлять, дополнять и уточнять взялся ван Занден — крупнейший голландский экономический историк. То есть база продолжает жить своей достаточно насыщенной жизнью, при том что база Мэддисона — это пример широко востребованного ресурса.

Наибольшую сложность представляет устойчивость информационных ресурсов, имеющих динамический характер содержащихся в них данных. При этом самостоятельную проблему составляет различение динамических и статических ресурсов. Кстати, не очевидно, что это различение строго дихотомично, возможны и промежуточные варианты.

Применительно к цифровой истории эта проблема рассмотрена в работе[1]. В более широком контексте вопросы специфики использования цифровых данных в исторических исследованиях в условиях цифрового поворота обсуждаются в работе[2].

Проблема устойчивости повторно используемых ресурсов имеет много аспектов. Можно начать с технологии представления динамичного научного знания в цифровой форме. Широко известна технология вики, предполагающая коллективную и анонимную работу над изменяемым текстом. Существует и альтернативная концепция, предполагающая сохранение авторства «живых документов»[3].

Существенное значение приобретает проблема идентификации динамичных информационных ресурсов и/или документов. Известно несколько подходов к идентификации изменяемых ресурсов и/или документов. Так, например, для программных инструментов, стандартов или классификационных языков часто применяется версионный подход. Существует подход, принятый в традиционном интернете, где ресурс идентифицируется местом размещения (URL). Возможна идентификация по времени фиксации ресурса. Однако все эти подходы имеют известные недостатки и часто критикуются.

---

[1] Клаверт Ф., Фикерс А. (2022) Публикация стипендии по цифровой истории в эпоху обновления. Журнал цифровой истории, 2 (1). https://doi.org/10.1515/JDH-2022–0003?locatt=label: JDHFULL

[2] Володин А.Ю. Шифры цифры: поиск ответов на трудные вопросы // Историческая информатика. 2019. № 3 (29). URL: https://cyberleninka.ru/article/n/shifry-tsifry-poisk-otvetov-na-trudnye-voprosy (дата обращения: 21.03.2023).

[3] Паринов С.И., Когаловский М.Р. Технология поддержки электронных научных публикаций как «живых» документов // Труды XI Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Петрозаводск: Карельский научный центр РАН, 17–21 сентября 2009. С. 53–58.

Многоуровневый характер в цифровой среде приобрела идентификация библиографии (концептуальные модели FRBR, LRM, BIBFRAME).

Для историков эта проблема имеет еще и источниковедческий аспект.

В литературе часто обсуждается проблема сопоставимости и достоверности данных, представленных в цифровых ресурсах. Обычно дискуссии вызывают методики сбора и представления данных. Во многих гуманитарных дисциплинах решение этой проблемы реализуется при помощи развитой системы метаданных цифровых ресурсов. В качестве примера можно привести фундаментальный подход к разработке метаданных для лингвистических ресурсов[1].

Технологически и организационно проблема включения цифровых ресурсов в инфраструктурные сервисы решается при помощи процедур сертификации в репозиториях цифровых данных. Ниже приводятся проекты по этой проблеме, реализованные в Евросоюзе. В российской цифровой гуманитаристике в качестве примера можно привести Репозиторий открытых данных по русской литературе и фольклору в Пушкинском доме (ИРЛИ)[2], в котором предусмотрена подобная процедура.

Однако вопросы сертификации для широкого класса информационных ресурсов по цифровой гуманитаристике нуждаются в широком обсуждении и общественном консенсусе. Тем более это относится к программным инструментам DH, разработка методов оценки которых практически не начата.

[1] Technologies for the Multilingual European Information Society. Specification of metadata-based descriptions for language resources and technologies/ Penny Labropoulou, Maria Gavrilidou, Elina Desipri, Stelios, Piperidis (R.C. Athena/ILSP), Francesca Frontini, Monica, Monachini (ILC/CNR), Victoria Arranz (ELDA), Gil Francopoulo (LIMSI) // Final Report, 2012. Режим доступа: http://www.meta-net.eu/public_documents/t4me/META-NET-D7.2.2-Final.pdf

[2] Репозиторий открытых данных по русской литературе и фольклору. URL: https://dataverse.pushdom.ru/ (доступно 20.03.2023).

Организация информационной инфраструктуры цифровых гуманитарных исследований в Европейском союзе

Прежде всего интересен опыт Евросоюза, где существует весьма развитая научная инфраструктура, ориентированная на развитие открытой науки и, конечно, широкое использование цифровых технологий, сервисов и коммуникаций. Достаточно подробное описание информационной инфраструктуры для социально-гуманитарных наук Евросоюза имеется в работе[1].

В настоящей статье мы приведем краткое описание институций, проектов и сервисов, имеющих прямое отношение к цифровой гуманитаристике.

Для создания и поддержания научной инфраструктуры в ЕС была создана специфическая организационно-правовая форма институции — Европейские консорциумы научной инфраструктуры (European Research Infrastructure Consortium — ERIC), которые предоставляют исследовательским сообществам ресурсы и услуги для проведения исследований и стимулирования инноваций.

Из 18 действующих в настоящее время ERIC некоторые были созданы специально для поддержки исследований в области цифровой гуманитаристики. Прежде всего это Цифровая научная инфраструктура для искусства и гуманитарных наук (DARIAH)[2], членами которой являются страны Европейского союза и ассоциированные государства и межправительственные организации. В настоящее время в DARIAH входят 20 стран-членов, 1 страна-наблюдатель и 6 стран-партнеров, включая Великобританию и США. В составе DARIAH создано несколько региональных центров. В ноябре 2022 года генеральная ассамблея DARIAH назначила директора DARIAH Агиатису Бенарду из Института систем управления информацией (ATHENA RC) и постдокторанта факультета информатики Афинского университета экономики и бизнеса.

DARIAH развивает и поддерживает цифровые исследования в области гуманитарных наук и искусства. Члены этой организации

---

[1] Антопольский А.Б. Информационная инфраструктура социально-гуманитарных наук в Евросоюзе // Наука и научная информация. 2021. № 4 (1–2). С. 8–22. URL: https://doi.org/10.24108/2658–3143–2021–4–1–2–18–32

[2] DARIAH — Digital Research Infrastructure for the Arts and Humanities European Research Infrastructure Consortium. URL: https://www.dariah.eu/ (дата обращения: 18.11.2022).

могут вносить свой вклад в сообщество в форме ресурсов, услуг и мероприятий. Вклад членов учитывается с помощью самооценки и экспертной оценки, которая проверяет совместимость вклада с инфраструктурой. Инструмент экспертной оценки используется для сбора, оценки и мониторинга национальных вкладов.

Прямое отношение к цифровой гуманитаристике имеет также Европейская научная инфраструктура языковых ресурсов и технологий (CLARIN ERIC)[1], ее деятельность описана в работе[2].

Другая инфраструктурная организация в этой области, недавно получившая статус ERIC, — это E-RIHS[3], — европейская исследовательская инфраструктура для науки о наследии, которая поддерживает исследования в области интерпретации, сохранения, документирования и управления наследием. Миссия E-RIHS заключается в предоставлении интегрированного доступа к экспертным знаниям, данным и технологиям на основе стандартизированного подхода, а также в интеграции ведущих европейских организаций и повышении их роли в научном сообществе глобального наследия.

E-RIHS ведет свою историю от проекта «Ариадна», направленного на интеграцию европейских исследовательских инфраструктур по археологическим наборам данных. E-RIHS имеет национальные структуры, или узлы, во многих странах — Великобритании, Греции, Италии, Швеции и др.

Кроме организаций и проектов, непосредственно относящихся к цифровой гуманитаристике, в европейском исследовательском пространстве реализовано множество более общих проектов и институций, образующих инфраструктуру социальных и гуманитарных наук в целом. Так, более широкие задачи ставит *OPERAS*[4] — открытая научная коммуникация в социальных и гуманитарных науках. Это исследовательская инфраструктура, поддерживающая открытое научное общение в области социальных и гуманитарных наук (SSH) в Европейском исследовательском пространстве. Его миссия состоит в том, чтобы координировать и объединять ресурсы в Европе для

[1] CLARIN ERIC — European Research Infrastructure for Language Resources and Technology. URL: https://www.clarin.eu/ (дата обращения: 18.11.2022).

[2] Антопольский А. Б. Лингвистические информационные ресурсы / науч. ред. Д. В. Ефременко. М.: ИНИОН РАН, 2022. 466 с.

[3] E-RIHS — European Research Infrastructure for Heritage Science. URL: http://www.e-rihs.eu/ (дата обращения: 18.11.2022).

[4] OPERAS — Открытые научные коммуникации в социальных и гуманитарных науках. URL: https://operas.hypotheses.org (доступно 18.11.2021).

эффективного удовлетворения научных потребностей европейских исследователей в этой области.

Отметим, что среди многочисленных инфраструктурных институций и проектов имеется специальная организация, занимающаяся долговременной устойчивостью инфраструктурных институций, ресурсов и сервисов[1]. Эта проблема особенно актуальна для России, где множество ресурсов и сервисов после прекращения финансирования, например грантового, перестают поддерживаться.

Направления и проекты развития инфраструктуры цифровой гуманитаристики

Перечислим инфраструктурные проекты, продукты и сервисы в области цифровой гуманитаристики, реализованные в различных институциях и проектах ЕС. Эти проекты сгруппированы по основным направлениям инфраструктуры.

Общие проекты

OpenAIRE[2] — единая точка входа к научным результатам в области цифровых гуманитарных наук и культурного наследия. В рамках этого проекта производится сбор результатов исследований, данных, научных публикаций и программных продуктов, относящихся к области цифровых гуманитарных наук. Это широкое определение включает гуманитарные науки, культурное наследие, историю, археологию и смежные области. В настоящее время в доступе находится около 4 млн публикаций, 300 тыс. наборов данных, свыше 800 программных средств.

SSHOC[3] — проект по представлению социальных и гуманитарных наук в Европейском облаке открытой науки (Облако SSH), где данные, инструменты и методики доступны для пользователей.

---

[1] Long-term sustainability. URL: https://research-and-innovation.ec.europa.eu/strategy/strategy-2020–2024/our-digital-future/european-research-infrastructures/long-term-sustainability_en (дата обращения: 18.11.2022).

[2] OpenAIRE Community Gateway for Digital Humanities and Cultural Heritage. URL: https://marketplace.eosc-portal.eu/services/digital-humanities-and-cultural-heritage-openaire-community-gateway/information

[3] SSHOC — Social Sciences & Humanities Open Cloud. URL: https://sshopencloud.eu/ (дата обращения: 18.11.2022).

Проект направлен на дальнейшее развитие инноваций, инфраструктурную поддержку цифровых методов в социальных и гуманитарных науках, развитие междисциплинарного сотрудничества, а также повышение воздействия SSH на общество.

PARTHENOS[1] направлен на укрепление сплоченности разработок в широком секторе лингвистических исследований, гуманитарных наук, культурного наследия, истории, археологии и смежных областях посредством тематического кластера европейских исследовательских инфраструктур, интеграции инициатив, электронных и других инфраструктур мирового класса.

CENDARI[2] — совместная Европейская цифровая архивная инфраструктура обеспечивает и облегчает доступ к существующим архивам и ресурсам в Европе для изучения средневековой и современной европейской истории посредством создания «среды исследований», которая расширяет подход к историческим записям по всему европейскому исследовательскому пространству, создав мощную новую платформу для доступа и исследования исторических данных транснациональным образом, преодолевая существующие в настоящее время национальные и другие хранилища данных.

SERISS[3] — проект «Интеграция европейских научных инфраструктур по социальным наукам» объединил все европейские консорциумы научной инфраструктуры в области социальных и гуманитарных наук. Его цель состояла в том, чтобы определить области возможного синергизма в развитии инфраструктуры и разработать ряд конкретных совместных мероприятий. Логическое обоснование этой идеи состояло в том, что двойные разработки следует предотвращать, инициативы должны взаимно извлекать выгоду из передовой работы других с целью создания совместных интегрированных доменов, где это имеет смысл для пользователей.

RISCAPE[4] — проект «Европейские исследовательские инфраструктуры в международном контексте». Его цель — представить аналитический отчет о положении и взаимодополняемости основных

[1] PARTHENOS — Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies. URL: http://www.parthenos-project.eu/about-the-project-2 (дата обращения: 18.11.2022).

[2] CENDARI — Collaborative European Digital Archival Research Infrastructure. URL: http://www.cendari.eu/ (дата обращения: 18.11.2022).

[3] SERISS — Synergies for Europe's Research Infrastructures in the Social Sciences. URL: https://seriss.eu/ (дата обращения: 18.11.2022).

[4] RISCAPE — European Research Infrastructures in the International Landscape. URL: https://riscape.wordpress.com/ (дата обращения: 18.11.2022).

европейских научных инфраструктур в контексте международной научной инфраструктуры. Отдельные разделы проекта посвящены социальным и гуманитарным наукам.

Публикации и репозитории

DARIAH-DE Working Papers[1] — сервис является центральным местом публикации различного рода материалов, созданных в контексте цифровых исследований гуманитарных наук. Они редактируются, но не подлежат никакому формальному процессу «экспертной оценки». Все материалы к рабочим документам DARIAH-DE публикуются в открытом доступе с лицензией CC-BY.

Hypotheses[2] — платформа по гуманитарным и социальным наукам, где размещено несколько тысяч блогов. Тексты находятся в открытом доступе и на разных языках. Hypotheses является частью более крупного портала OpenEdition, который представляет собой комплексную цифровую издательскую инфраструктуру для распространения публикаций по гуманитарным и социальным наукам.

DARIAH-DE Repository[3] является центральным компонентом архитектуры DARIAH-DE. Репозиторий позволяет хранить исследовательские данные устойчиво и надежно, предоставлять пользователям метаданные и находить их с помощью универсального поиска. Пользователь может импортировать свои исследовательские данные с помощью DARIAH-DE Publikator[4].

ARCHLAB[5] предоставляет доступ к организованной научной информации в форме в основном неопубликованных наборов данных из архивов престижных европейских музеев, галерей и исследовательских институтов. ARCHLAB обеспечивает доступ к объединенным знаниям в 14 хранилищах в Бельгии, Великобритании, Германии, Испании, Италии, Нидерландах, Румынии, Франции и Швеции.

---

[1] DARIAH-DE Working Papers. URL: https://de.dariah.eu/en/working-papers (дата обращения: 18.11.2022).

[2] Hypotheses. A platform for humanities and social science research blogs. URL: https://hypotheses.org/ (дата обращения: 18.11.2022).

[3] DARIAH-DE Repository. URL: https://de.dariah.eu/en/repository (дата обращения: 18.11.2022).

[4] DARIAH-DE Publikator. URL: https://de.dariah.eu/en/publikator (дата обращения: 18.11.2022).

[5] ARCHLAB. URL: https://www.iperionhs.eu/archlab/ (дата обращения: 18.11.2022).

DARIAH at HAL[1] — открытый архив HAL, в который авторы помещают научные тексты по тематике всех академических областей. Услуга доступна на английском и французском языках. Это вклад в DARIAH с 2015 года французского Centre pour la Communication Scientifique Directe (CCSD).

TextGrid[2] — виртуальная исследовательская среда, которая позволяет ученым-гуманитариям редактировать и публиковать свои цифровые документы и данные, а также надежно их хранить.

Платформы

ISIDORE[3] — платформа и поисковая система, предоставляющая доступ к цифровым публикациям и данным гуманитарных и социальных наук. Открытая для всех и особенно для преподавателей, исследователей, аспирантов и студентов, она опирается на принципы семантической сети и обеспечивает открытый доступ к данным. *ISIDORE* предлагает более 5 млн ресурсов всего мира, а обогащения доступны на трех языках: французском, английском и испанском.

IPERION HS[4] — интегрирующая платформа для европейской исследовательской инфраструктуры в области науки о наследии — это консорциум из 24 национальных узлов в 23 странах, в том числе 67 организаций. Он предлагает обучение и доступ к широкому спектру научных инструментов, методологий, данных и инструментов высокого уровня для продвижения знаний и инноваций в области науки о наследии.

DARIAH Wiki[5] — платформа для исследовательских проектов, связанных с DARIAH. Проекты и исследовательские группы могут получить свое собственное пространство Вики.

ConedaKOR[6] — веб-система баз данных с графовой архитектурой, предназначенная для администрирования и презентации академических материалов и коллекций объектов из области изобразительного искусства, культурологии и гуманитарных наук. Система позволяет хранить произвольные документы и соединять

[1] DARIAH at HAL. URL: https://hal.archives-ouvertes.fr/DARIAH (дата обращения: 18.11.2022).

[2] TextGrid. URL: https://textgrid.de/en/ (дата обращения: 18.11.2022).

[3] ISIDORE. URL: https://isidore.science/ (дата обращения: 18.11.2022).

[4] IPERION HS. URL: https://www.iperionhs.eu/ (дата обращения: 18.11.2022).

[5] DARIAH Wiki. URL: https://wiki.de.dariah.eu/ (дата обращения: 18.11.2022).

[6] ConedaKOR. URL: https://github.com/coneda/kor#readme (дата обращения: 18.11.2022).

их с другими документами через отношения, строить огромные семантические сети из неограниченного количества доменов.

Библиография и каталоги

Doing Digital Humanities[1] — «Создавая цифровую гуманитаристику» — постоянная деятельность по сбору библиографических описаний текстов, касающихся цифровых гуманитарных проблем. Для сервиса используется Zotero — бесплатный и простой в использовании инструмент, который помогает собирать, систематизировать, цитировать и делиться своими исследовательскими источниками.

Collection Registry[2] — реестр коллекций, простое веб-приложение, содержащее информацию о научных коллекциях, имеющих отношение к социальным и гуманитарным исследованиям.

Словари и онтологии

Vocabs[3] — контролируемые словари (справочники, тезаурусы и т.д.) обеспечивают качество ресурсов и взаимодействие между ними во многих областях научной деятельности. Сервис предоставляет услуги и инструменты, которые позволяют совместно создавать, поддерживать и публиковать словари и таксономии любого рода. Система основана на программном обеспечении с открытым исходным кодом Skosmos, которое использует SKOS в качестве базовой модели данных.

TaDiRAH[4] — таксономия цифровой исследовательской деятельности в гуманитарных науках разработана для использования общественными сайтами и проектами, направленными на структурирование информации, относящейся к цифровым гуманитарным наукам. Отдельные фасеты таксономии — это виды деятельности, научные объекты, научные методики.

---

[1] Doing Digital Humanities. URL: https://www.zotero.org/groups/113737/doing_digital_humanities_-_a_dariah_bibliography (дата обращения: 18.11.2022).

[2] Collection Registry. URL: https://colreg.de.dariah.eu/colreg/?lang=en (дата обращения: 18.11.2022).

[3] Vocabs. URL: https://vocabs.dariah.eu/en/ (дата обращения: 18.11.2022).

[4] TaDiRAH — Taxonomy of Digital Research Activities in the Humanities. URL: http://tadirah.dariah.eu/vocab/index.php, https://vocabs.dariah.eu/tadirah2/en/ (дата обращения: 18.11.2022).

NeMO[1] — это комплексная онтологическая модель научной практики в области социальных и гуманитарных наук, разработка которой осуществляется через исследовательскую сеть ESF NeDiMAH[2]. *NeMO* — это CIDOC CRM-совместимая онтология, которая явно опирается на факторы агентов (акторов), процессов (деятельности и методов) и ресурсов (информационных ресурсов, инструментов, концепций), проявляющихся в научном процессе. Онтология основана на результатах обширных эмпирических исследований и моделирования научных практик, выполненных в проектах DARIAH.

Образовательные ресурсы

DARIAH-Campus[3] — это одновременно платформа обнаружения и хостинговая платформа для предложений в области обучения и образования.

PARTHENOS Training Suite[4] — учебный комплекс, разработанный на основе проекта PARTHENOS, объединившего экспертные знания в области искусства, гуманитарных и социальных наук по управлению данными, стандартам и обучению, которые должны быть размещены на специально построенной платформе.

Digital Humanities Course Registry[5] — открытый онлайн-реестр модулей, курсов и программ DH в Европе.

Online platform for teaching Digital Humanities (#dariahTeach)[6] — открытые, управляемые сообществом, многоязычные высококачественные учебные материалы для цифровых методов в искусстве, социальных и гуманитарных науках.

[1] NeMO — NeDiMAH Methods Ontology. URL: http://nemo.dcu.gr (дата обращения: 18.11.2022).

[2] NeDiMAH — Network for Digital Methods in the Arts and Humanities. URL: http://archives.esf.org/coordinating-research/research-networking-programmes/humanities-um/nedimah.html (дата обращения: 18.11.2022).

[3] DARIAH-Campus. URL: https://campus.dariah.eu/ (дата обращения: 18.11.2022).

[4] PARTHENOS Training Suite. URL: http://training.parthenos-project.eu/ (дата обращения: 18.11.2022).

[5] Digital Humanities Course Registry. URL: https://dhcr.clarin-dariah.eu/courses (дата обращения: 18.11.2022).

[6] Online platform for teaching Digital Humanities. URL: https://www.youtube.com/channel/ UCScSbG7XjiXbZVgilEp0Pkw (дата обращения: 18.11.2022).

Программные инструменты

Generic Search[1] — представляет собой распределенный метапоиск по коллекциям и ресурсам, хранящимся в реестрах DARIAH, независимо от их схем данных и метаданных. Он устанавливает и отслеживает семантические связи между структурно различными коллекциями и их конкретными ресурсами.

DARIAH docs[2] — сервис, представляющий альтернативу *Google Docs*, что очень востребовано научным сообществом. Он размещается и управляется компанией *DAASI International*[3] и основан на *Collabora Online*. После входа в систему можно создавать совместные документы.

HedgeDoc (ранее *CodiMD)*[4] — онлайн-инструмент с открытым исходным кодом, который позволяет нескольким пользователям работать с одним текстом одновременно из разных мест.

Geo-Browser[5] — поддерживает визуализацию данных в привязке к географическим пространственных данным и к определенным периодам времени; исследователи могут анализировать пространственно-временные отношения данных и коллекций и строить корреляции между ними.

MEISE[6] — редактор партитур MEI — программное обеспечение для редактирования нотной записи для корректуры и редактирования музыки, записанной в CWN (Common Western Notation), а также для визуализации вариантов и прочтений.

Лингвистические процессоры

DARIAH-DE Topics Explorer[7] — метод анализа распределения семантических кластеров слов, так называемых «тем», в текстовой коллекции. Он может быть использован для изучения содержимого корпуса, а также для генерирования связанных с ним признаков для

---

[1] Generic Search. URL: https://search.de.dariah.eu/search/?lang=en (дата обращения: 18.11.2022).

[2] DARIAH docs. URL: https://daasi.de/en/digital-humanities-english/dariahdocs/ (дата обращения: 18.11.2022).

[3] DAASI International. URL: https://daasi.de/en/ (дата обращения: 18.11.2022).

[4] HedgeDoc. URL: https://pad.gwdg.de/ (дата обращения: 18.11.2022).

[5] Geo-Browser. URL: https://geobrowser.de.dariah.eu/ (дата обращения: 18.11.2022).

[6] DARIAH-DE MEI Score Editor Webservice. URL: http://meise.de.dariah.eu/ (дата обращения: 18.11.2022).

[7] DARIAH-DE Topics Explorer. URL: https://de.dariah.eu/en/web/guest/topicsexplorer (дата обращения: 18.11.2022).

классификации цифрового текста. Тематическое моделирование полностью опирается на анализируемые тексты; оно не использует дополнительных источников информации, таких как словари или внешние обучающие данные, что делает его в значительной степени независимым от языка и орфографических условностей.

DKPro Wrapper[1] — оболочка для программного продукта компании DKPro, ориентированная на извлечение лингвистической информации из книг. В руководстве пользователя объясняется простое управление этим инструментом при помощи командной строки на Java.

Стандартизация

Standardization Survival Kit (SSK)[2] — сервис для поддержки цифровых методов в социальных и гуманитарных науках, где необходимы знания о стандартах и передовой исследовательской практике. Цель сервиса, разрабатываемого в рамках проекта *PARTHENOS*, заключается в том, чтобы предоставлять исследователям доступ к стандартам и передовой практике на всех этапах и при всех методиках исследований. SSK — это открытый инструмент, в котором пользователи могут публиковать новые методики или адаптировать существующие.

Идентификация

AAI[3] — инфраструктура аутентификации и авторизации (DARIAH AAI) обеспечивает федеративный единый вход, позволяющий использовать несколько сервисов через одну учетную запись.

---

[1] DARIAH-DKPro-Wrapper v0.4.7. URL: http://dariah-de.github.io/DARIAH-DKPro-Wrapper/user-guide.html (дата обращения: 18.11.2022).

[2] Standardization Survival Kit (SSK). URL: https://www.dariah.eu/tools-services/tools-and-services/tools/standardization-survival-kit/ (дата обращения: 18.11.2022).

[3] The DARIAH Authentication and Authorization Infrastructure (DARIAH AAI). URL: https://wiki.de.dariah.eu/display/publicde/DARIAH+AAI+Documentation (дата обращения: 18.11.2022).

Мероприятия

DARIAH Calenda[1] — календарь гуманитарных и социальных наук — это онлайн-сервис объявлений с открытым доступом в области гуманитарных и социальных наук. Он информирует студентов, преподавателей и исследователей о текущем состоянии исследований.

Коммуникация

DHd-Blog[2] — цифровая гуманитаристика в немецкоязычных странах. Цель веб-сайта и блога — максимально широко представить DH-исследовательские сообщества Австрии, Германии и Швейцарии.

Аппаратура и оборудование

FIXLAB[3] — доступ сообщества Heritage Science (HS) к ключевым стационарным исследовательским установкам и связанному с ними научному опыту их сотрудников, которые разрабатывают и поддерживают сложные современные приборы для диагностики и археометрии.

MOLAB[4] (мобильная лаборатория) — это распределенная инфраструктура мирового уровня, состоящая из ключевых лабораторий в 10 европейских странах, обеспечивающая согласованный доступ под единой структурой управления к набору мобильного оборудования и смежных компетенций для неразрушающих измерений произведений искусства, коллекций, памятников и объектов.

Приведенный перечень инфраструктурных проектов, продуктов и сервисов в области цифровой гуманитаристики дает полное представление о широте и размахе их реализации в Евросоюзе. Подробный обзор инфраструктурных проектов по компьютерной лингвистике, которую часто включают в цифровую гуманитаристику, имеется также в монографии[5]. Можно добавить, что подобные

[1] DARIAH Calenda. URL: https://calenda.org/search.html?q=dariah (дата обращения: 18.11.2022).

[2] DHd-Blog — Digital Humanities im deutschsprahigen Raum. URL: https://dhd-blog.org/ (дата обращения: 18.11.2022).

[3] FIXLAB. URL: https://www.iperionhs.eu/fixlab/ (дата обращения: 18.11.2022).

[4] MOLAB (Mobile LABoratory). URL: https://www.iperionhs.eu/molab/ (дата обращения: 18.11.2022).

[5] Клаверт Ф. и Фикерс А. (2022) Публикация стипендии по цифровой истории в эпоху обновления. Журнал цифровой истории, 2 (1). https://doi.org/10.1515/JDH-2022–0003?locatt=label: JDHFULL

инфраструктурные проекты, ресурсы и сервисы реализуются также в США, Канаде, Японии и других странах.

Состояние российской информационной инфраструктуры

На фоне состояния научной информационной инфраструктуры Европы российские достижения выглядят достаточно скромно. Прежде всего не хватает общей концепции развития инфраструктуры в условиях цифровизации науки. Хотя существует множество различных органов по управлению наукой и высокими технологиями (Министерство науки и высшего образования, Министерство цифрового развития, связи и массовых коммуникаций, Российская академия наук, Российский научный фонд, разнообразные научные советы, в том числе при Президенте), в стране не выработано ясного представления о необходимой инфраструктуре для поддержки цифровизации науки, которая опиралась бы на реальное состояние дел.

Напомним, что в 2018–2019 годах были разработаны и размещены на сайте Минобрнауки *Концепция цифровой автоматизированной системы предоставления сервисов научной инфраструктуры коллективного пользования* (АС УСНИКП) и *Концепция создания Единой цифровой платформы научного и научно-технического взаимодействия, организации и проведения совместных исследований в удаленном доступе, в том числе с участием зарубежных ученых* (ЦПСИ). Однако эти концепции не были реализованы.

В 2021 году появился новый документ «Стратегия цифровой трансформации отрасли науки и высшего образования»[1]. Отметим его особенности. Во-первых, он в основном ориентирован на цифровизацию высшего образования, и специфика информационного обеспечения фундаментальных исследований в нем практически не учитывается. Во-вторых, в документе полностью отсутствует анализ состояния научных информационных ресурсов России, на который, как представляется, должна опираться стратегия. В-третьих, предлагаемый для управления данными проект «Датахаб» никак не связан с существующими банками научных данных. Складывается

[1] Стратегия цифровой трансформации отрасли науки и высшего образования. М.: Минобрнауки, 2021. URL: https://minobrnauki.gov.ru/upload/iblock/e16/ dv6edzmr0og5dm57dtm0wyllr6uwtujw.pdf

впечатление, что разработчики имели в виду только отчетные управленческие данные, а не результаты научных исследований, хотя прямо в документе об этом не сказано. В проекте вообще не упоминаются электронные библиотеки, архивы и репозитории научных данных, которые в последние годы становятся основными хранилищами научной информации.

Можно предположить, что существенным компонентом научной инфраструктуры может стать Российский центр научной информации (РЦНИ), устав которого недавно был утвержден Постановлением Правительства РФ № 1357 от 29 июля 2022 г. «О федеральном государственном бюджетном учреждении "Российский центр научной информации"» (https://www.fbras.ru/wp-content/uploads/2022/08/Postanovlenie-RF-----1357-ot-29-iyulya-2022-goda.pdf). В уставе, в частности, указано, что РЦНИ осуществляет методологическую поддержку управления научными данными, международное сотрудничество в научно-информационной сфере, создание и поддержку информационно-аналитических систем, баз данных и цифровых платформ, экспертизу проектов в информационной сфере и другие инфраструктурные функции.

Однако пока РЦНИ реализует только проекты, связанные с научной периодикой, — создает платформу для электронных научных журналов, участвует в создании «белого списка» научных журналов, а также занимается подпиской на зарубежные ресурсы. Другие инфраструктурные проекты РЦНИ нам неизвестны, хотя в предварительных документах среди функций этого центра упоминались, в частности, служба идентификации, а также поддержка классификаций.

Важным недостатком существующих проектов и директив представляется отсутствие стимулов для научной коллаборации и создания коллективных цифровых научных проектов, которые могли бы стать инструментом для повышения эффективности научных исследований. В качестве примера успешного российского научного проекта такого рода можно назвать Национальный корпус русского языка[1], разработанный под руководством акад. В. А. Плунгяна. Этот проект создал принципиально новые возможности не только для фундаментальных лингвистических исследований, но и для решения многих прикладных задач.

[1] Национальный корпус русского языка. URL: https://ruscorpora.ru/ (дата обращения: 18.11.2022).

Из перспективных, но нереализованных российских инициатив в области научной информационной инфраструктуры следует упомянуть идею Единого российского электронного пространства знаний (ЕРЭПЗ), изложенную в Постановлении Правительства РФ от 20.02.2019 № 169 «Об утверждении Положения о федеральной государственной информационной системе "Национальная электронная библиотека" и методики отбора объектов Национальной электронной библиотеки» (https://www.garant.ru/products/ipo/prime/doc/72084144/). По этой проблеме было проведено несколько интересных исследований, однако реалистичной программы создания ЕРЭПЗ не было предложено. Значительные ассигнования были выделены для создания российской электронной энциклопедии на портале «Знание», однако этот проект оказался никак не связан с ЕРЭПЗ. Подробно эти вопросы рассмотрены в монографии[1].

К инфраструктурным сервисам следует отнести службу поддержки классификаторов научной информации (ГРНТИ, УДК). Ранее она действовала при ВИНИТИ РАН, а в последние годы передана в ГПНТБ России. Однако результаты деятельности этой службы существенно отстают от потребностей современной науки. Критика современного состояния ГРНТИ содержится в работе[2].

Можно упомянуть также отдельные российские исследования по семантической сети и открытым связанным данным, по созданию онтологий, а также навигаторов информационных ресурсов. Эти разработки могли бы стать компонентами российской научной информационной инфраструктуры, если бы существовала программа и координация деятельности в этой области.

Задачи создания российской инфраструктуры для цифровой гуманитаристики

С учетом международного опыта и актуальных потребностей российской научной инфосферы программа создания научной

[1] Антопольский А.Б. Научная монография и электронное пространство знаний / науч. ред. Д.В. Ефременко. М.: ИНИОН РАН, 2020. 252 с.

[2] Антопольский А.Б. Языки индексирования для цифровой гуманитаристики // Научно-техническая информация. Сер. 2. 2022. № 1. С. 1–9. DOI: 10.36535/0548–0027–2022–01–1.

инфраструктуры цифровой науки могла бы включать следующие основные направления:

- мониторинг и учет разрабатываемых информационных ресурсов, ведение официальных государственных каталогов;
- создание и поддержка репозиториев научных данных и электронных библиотек;
- сертификация цифровых информационных ресурсов и программных инструментов;
- наукометрические исследования и измерения;
- интеграция информационных ресурсов, прежде всего на основе платформы семантической сети и открытых связанных данных;
- архивирование и долговременное хранение сертифицированных информационных ресурсов, обеспечивающих их повторное использование;
- идентификация информационных объектов;
- образовательные программы по обучению прогрессивным технологиям и обмену передовым опытом;
- поддержка стандартов, методик, технологий, программных инструментов и систем метаданных, авторитетных файлов, тезаурусов и других средств лингвистического обеспечения.

Необходимо добавить, что такая программа должна учитывать особенности и потребности основных направлений фундаментальной и прикладной науки, а именно: естественные, социо-гуманитарные, технические, медицинские и сельскохозяйственные науки.

Легко показать, что каждое из этих направлений нуждается в специфических инфраструктурных проектах, ресурсах и сервисах, в том числе метаданных, онтологий, каталогах и т.д. К тому же в каждом из этих направлений исторически сложились головные организации и соответствующие сообщества.

Одним из компонентов будущей инфраструктуры российской цифровой гуманитаристики должна стать Справочно-информационная система по цифровой гуманитаристике, упомянутая выше и описанная в работе[1].

[1] Антопольский А. Б., Володин А. Ю. Справочно-информационная система по цифровой гуманитаристике: опыт описания интернет-ресурсов российских архивов // Историческая информатика. 2022. № 2. С. 50–66. DOI: 10.7256/2585–7797.2022.2.38236 EDN: HOVPGY URL: https://nbpublish.com/library_read_article.php?id=38236

Послесловие

Отдайте же человеку — человеческое, а вычислительной машине — машинное.

Норберт Винер

Даже самый беглый обзор цифровых гуманитарных наук позволяет понять, что компьютеры открыли перед учеными пространство больших возможностей. Слово «большой» в контексте разговоров о настоящем и будущем науки не случайно. В зависимости от угла зрения, наше время называют и эпохой больших данных, и эпохой больших языковых моделей. Большими они называются не только ради рекламной гиперболы, но и потому, что эти сущности больше, чем каждый человек в отдельности.

Традиционный образ ученого-гуманитария рисует нам одиночку, который, запершись в кабинете наедине с текстом, погружен в герменевтику или метафизику.

Большие данные и компьютерные методы их анализа, во-первых, возвращают гуманитария к эмпирическому материалу, позволяют ему нарисовать перед собой большую картину знания, масштаб которой раньше был недоступен; во-вторых, позволяют ученому преодолеть конечность собственных исследовательских возможностей; и, в-третьих, позволяют объединять в научном диалоге многих специалистов, становящихся участниками общего проекта.

Благодаря новой конфигурации научных коллективов, благодаря новым исследовательским вопросам и электронным копиям недоступных ранее объектов изучения меняется сам портрет гуманитария.

Если ни один человек не может прочесть за свою жизнь миллионы томов, то компьютер — может. Если ни один человек не может собрать в своем кабинете полную коллекцию археологических находок, то электронное хранилище — может. Содружество гуманитария и компьютера не только позволяет развернуть исследовательские направления, которые раньше представляли собой чистую фантазию,

но и заставляет эволюционировать стоящего в центре этой сферы субъекта.

Если еще несколько десятилетий назад имели право на публикацию работы, в которых были выполнены только какие-то подсчеты, то теперь количественные характеристики без интерпретаций публиковать неприлично. Считать умеют уже все. А вот объяснять результаты подсчетов, извлекать из них новое знание способны только специально подготовленные к таким задачам исследователи.

Цифровые исследования для гуманитария — это окно из его кабинета в большой мир, такой, который до появления компьютерных помощников трудно было охватить даже мысленным взором. Теперь охватить можно гораздо больше, но и научную оптику (тот самый «взор») нужно перенастраивать. Недаром возникают такие понятия цифровой гуманитарной оптики, как, например, «дальнее чтение» или «макроскоп». В нашей книге для этой перенастройки не всегда можно найти готовые инструкции, но вдумчивый читатель обязательно сможет нащупать направление, в котором ему следует двигаться.

Digital humanities все еще переживают процесс становления. Далеко не все проблемы решены, далеко не все пути найдены. Поэтому цифровые гуманитарные науки заинтересованы в привлечении в свои ряды свежих сил. Для этого нужно стараться рассказывать на всех языках потенциальным союзникам о возможностях и ограничениях доступных в рамках этой дисциплины методов, о достижениях и слабостях ведущихся исследований, о решаемых и малоперспективных научных вопросах. Именно такой диалог на русском языке мы и выстраиваем с читателем (и, как мы рассчитываем, в скором времени коллегой) на этих страницах. Книга на русском языке, но при этом она фиксирует мировые тренды и ссылки, которые в ней можно найти, чаще на иностранных языках. Этот принцип можно выразить с помощью англоязычного афоризма: think global, act local.

За последние годы уже сложился определенный «канон» компьютерных методов, который необходимо иметь в виду, обращаясь к машиночитаемым данным или рассчитывая увидеть смысл в оцифрованных коллекциях, превышающих возможности нашего физического восприятия. К такому «канону» можно отнести базы данных, компьютерный анализ текста, геоинформационный анализ, сетевой анализ данных, компьютерное моделирование. При этом в большинстве случаев эти методы — в исконном значении путь исследования, который каждый гуманитарий проходит по-своему,

собирая уникальное соотношение нужных подходов и умений, существенно обогащающих присущие исследователям навыки вдумчивого чтения, пристального наблюдения и интуитивной классификации. Чем более сложным станет инструментарий исследователя, тем сильнее он будет себе казаться. Но не стоит обольщаться, полагая, что методы заменят знания. Важно понимать, что цифровые методы и данные ни в коем случае не заменяют глубокого знания и понимания предмета исследования. Пути формализации и концептуализации могут быть разными, но главным остается приращение нового знания. Для того чтобы методы применить успешно, нужно хорошо знать предмет, который предполагается измерить, изучить, понять.

В книге мы не раз подчеркиваем, что цифровые гуманитарные исследования — это молодая и динамично развивающаяся область. В то же время мы наблюдаем, как пионеры этой сферы быстро становятся пенсионерами. Стало быть, сама дисциплина уже выходит из детского возраста. Ее состояние хорошо описывается цитатой из классического романа «Мельмот Скиталец» Ч. Р. Метьюрина, серьезно повлиявшего на «Евгения Онегина» и «Портрет Дориана Грея»: « …его нежное румяное лицо, стройная, словно точеная фигура и переливы его мягкого голоса пробуждали в людях тот смешанный интерес, с каким мы обычно наблюдаем, как в юноше сквозь еще отроческую незрелость пробиваются первые побеги силы, которым в будущем суждено вырасти и окрепнуть, и наполняли сердца родителей той ревнивой тревогой, с какой мы следим за погодой теплым, но сумрачным весенним утром: мы радуемся разлитому в небе спокойному сиянию зари, однако боимся, что еще до полудня лазурь его будет затянута тучами». И хотя до зрелости цифровым гуманитарным наукам еще далеко, сейчас она определенно находится в точке юношеской равновесности, в том моменте, когда еще рано подводить окончательные итоги, но уже нужно фиксировать промежуточные. Надеемся, что наша книга послужит и этому. Carpe diem, quam minimum credula postero.

Андрей Володин,
Борис Орехов

Информация об авторах

Антопольский Александр Борисович — доктор технических наук, главный научный сотрудник Института научной информации по общественным наукам РАН.

Бонч-Осмоловская Анастасия Александровна — кандидат филологических наук, доцент Школы лингвистики факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики», DH-Cloud.

Бородкин Леонид Иосифович — член-корреспондент РАН, доктор исторических наук, заведующий кафедрой исторической информатики исторического факультета Московского государственного университета имени М. В. Ломоносова.

Володин Андрей Юрьевич — кандидат исторических наук, доцент исторического факультета Московского государственного университета имени М. В. Ломоносова, руководитель стратегического проекта «Институт цифровых гуманитарных исследований» Сибирского федерального университета.

Гагарина Динара Амировна — кандидат педагогических наук, научный сотрудник Университета имени Фридриха — Александра в Эрлангене и Нюрнберге, DH-Cloud.

Гришин Евгений Сергеевич — руководитель сектора исторической картографии и геоинформационных систем Научно-исследовательской лаборатории экономической и социальной истории Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации.

Кижнер Инна Александровна — кандидат культурологии, доцент Гуманитарного института Сибирского федерального университета, научный сотрудник Элияху Лаборатории цифровых гуманитарных исследований Хайфского университета.

Орехов Борис Валерьевич — кандидат филологических наук, доцент Школы лингвистики факультета гуманитарных наук и ведущий научный сотрудник Международной лаборатории языковой конвергенции Национального исследовательского университета «Высшая школа экономики», старший научный сотрудник Лаборатории цифровых исследований литературы и фольклора Института русской литературы (Пушкинского дома) РАН.

Румянцев Максим Валерьевич — кандидат философских наук, ректор Сибирского федерального университета, научный руководитель лаборатории «Digital Humanities» Сибирского федерального университета.

Сметанин Андрей Владимирович — кандидат исторических наук, доцент историко-политологического факультета Пермского государственного национального исследовательского университета.

Научное издание

Антопольский Александр Борисович
Бонч-Осмоловская Анастасия Александровна
Бородкин Леонид Иосифович
Володин Андрей Юрьевич
Гагарина Динара Амировна
Гришин Евгений Сергеевич
Кижнер Инна Александровна
Орехов Борис Валерьевич
Румянцев Максим Валерьевич
Сметанин Андрей Владимирович

ЦИФРОВЫЕ ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ

Монография

Редактор Л. А. Киселева
Компьютерная верстка И. В. Гревцовой

Подписано в печать 20.09.2023. Печать плоская
Формат 60×84/16. Бумага офсетная. Усл. печ. л. 17,0
Тираж 500 экз. Заказ № 19926

Библиотечно-издательский комплекс
Сибирского федерального университета
660041, Красноярск, пр. Свободный, 82а
Тел. (391) 206-26-16; http://bik.sfu-kras.ru
E-mail: publishing_house@sfu-kras.ru

Для заметок

Для заметок