View metadata, citation and similar papers at core.ac.uk
brought to you by CORE

DOI: 10.21514/1998-426X-2017-10-4-44-52
УДК: 614.876:004.6

Проблемы аналитического обеспечения коммуникации рисков: обоснование подходов к разработке исследовательских баз данных по вопросам радиационной безопасности и социальных рисков

Л.С. Рехтина[1], Н.В. Соколов[1], А.М. Библин[2], Л.В. Репин[2], Р.Р. Ахматдинов[2]

[1]С:анкт-Петербургский государственный университет, Правительство Российской Федерации, Санкт-Петербург, Россия
[2]Санкт-Петербургский научно-исследовательский институт радиационной гигиены имени профессора П.В. Рамзаева Федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека, Санкт-Петербург, Россия

Одним из важных этапов риск-коммуникации является анализ публикаций в традиционных средствах массовой информации и сети Интернет, которые в значительной степени формируют отношение людей к различным проблемам. В то же время доступность большого количества информационных материалов, относящихся к любой проблемной области, затрудняет возможности ручного анализа и адекватного описания всего объема информации. С другой стороны, доступность информации обусловливает актуальность разработки методов повышения эффективности ее анализа. Одним из способов автоматизации анализа больших объемов информации является разработка баз данных или автоматизированных информационных систем, содержащих информационные материалы по изучаемой проблематике и предполагающих возможность автоматизированной обработки. Целью данной работы является анализ опыта разработки таких систем и баз данных научными коллективами Санкт-Петербургского научно-исследовательского института радиационной гигиены и Санкт-Петербургского государственного университета и выявление ключевых особенностей применения баз данных для социальных исследований. Результаты проведенного анализа показали, что использованные методические подходы очень близки. Анализ выполнен по методике автоэтнографического исследования. Применение стратегии сравнительного анализа позволяет выявить общие черты, характеризующие ситуацию разработки и внедрения информационных систем и баз данных в практику анализа информационного поля. В статье рассматриваются особенности и связанные с ними ограничения первичных данных, такие как: текстовый, дискурсивный характер большинства материалов, информационный шум, высокая зависимость от контекста, изменчивость, разная структура, формат и вид материалов. Приводятся значимые для решения задач качественного и количественного анализа параметры. Важным условием создания эффективной, с точки зрения социально-коммуникационных исследований информационной системы, является реализация возможности обработки собранных социальных данных с помощью доступных исследователям программ. Требование автоматизации ряда функций упирается в высокую степень уникальности материалов, которая пока преодолевается только ручной пре- и постобработкой, что необходимо учитывать при проектировании исследовательских программ и систем автоматизированной обработки информации.

Ключевые слова: *базы данных, автоматизированные информационные системы, радиационная безопасность, риск-коммуникация, анализ публикаций, социальная рискология, социология науки, социология миграции.*

Введение

В настоящей статье суммируется опыт разработки информационных систем и баз данных для решения задач автоматизации анализа описания социальных явлений в средствах массовой информации (СМИ) и социальных сетях, накопленный научными коллективами Санкт-Петербургского научно-исследовательского института (НИИ) радиационной гигиены имени профессора П.В. Рамзаева и Санкт-Петербургского государственного университета (СПбГУ) [1, 2]. Автоматизированные

Рехтина Лилия Сергеевна
Санкт-Петербургский государственный университет.
Адрес для переписки: 191124, Россия, Санкт-Петербург, ул. Смольного, д. 1/3, 9-й подъезд; E-mail: lisabet-09@mail.ru

информационные системы (АИС) и базы данных (БД) давно стали неотъемлемой частью всех сфер деятельности, в которых применяются информационные технологии [3], включая и социальные исследования, авторы которых все чаще признают: собранная и обработанная с помощью автоматизированных систем информация приобретает статус важнейшего исследовательского ресурса [4]. Эффективное применение АИС и БД для целей коммуникационного и социального анализа обладает рядом особенностей. Специфика объектов исследования, представляющей их информации, методов ее обработки и анализа создают так много препятствий для внедрения автоматизированных систем, что до сих пор продолжают доминировать «традиционные» (а на самом деле – морально устаревшие) методики. Например, большинство заказчиков отдают предпочтение привычным социологическим опросам, с недоверием относясь к возможностям исследований, базирующихся на анализе интернет-материалов.

Сфера коммуникации риска – одна из областей практической деятельности, где интернет-исследования особенно востребованы и происходит постепенное внедрение АИС и БД [5, 6]. Это объясняется способностью Интернета отражать, а иногда и выявлять общественную рефлексию различных рисков. Развитие Интернета неслучайно совпало во времени с формированием «общества риска» [7], которое само воспроизводит и даже приумножает риски различной природы. Например, в некоторых сферах жизни мифы, существующие и распространяющиеся в общественном сознании, имеют более устойчивый характер, чем объективные научные знания [1]. При этом общественностью проблематизируются объекты, риски от которых нормируются и контролируются, как официальными, так и социальными субъектами (например, объекты атомной отрасли) [8]. И, напротив, совершенно упускаются из виду потенциально более опасные виды радиационного воздействия (например, радон и медицинское облучение пациентов), которые как раз должны быть частью постоянного общественного обсуждения и социального контроля [2].

Процесс принятия решений в социально значимых отраслях (например, атомной) сопровождаются недостатком доступной информации, особенно в отношении вопросов, возведенных массовым сознанием в ранг проблемы. Проблемой, требующей внимательного отношения, становится нагнетание рисков и их мифологизация [9]. Проблемы поиска информации меняют свой ракурс – фокус смещается с поиска информации на поиск достоверной информации и ее верификацию, что зачастую требует обработки большого объема данных, превышающего возможности «ручной» обработки одним или даже несколькими специалистами [10]. Анализ и верификация информации в рассматриваемой нами области требуют уже не столько коллективного труда, сколько автоматизации анализа и обработки данных при решении прикладных задач. Исследования требуют постоянного наращивания объема массива интернет-данных [11] с разными структурными свойствами, а ключевой задачей управления данными становится проектирование БД и АИС, при котором особое внимание должно уделяться структуризации данных и разработке критериев оценки информационных материалов с различных прикладных точек зрения. Недостаточно просто собрать данные в одно большое «хранилище», необходимо предусмотреть алгоритмы поиска и систематизации, описания взаимосвязей между материалами, определить структуру, формат и форму их хранения с тем, чтобы они были пригодны для применения различных методов анализа и программ обработки данных.

Цель исследования – обоснование практических подходов к разработке АИС и БД, предназначенных для решения задач автоматизации анализа информационного поля в области риск-коммуникации, при реализации научными коллективами НИИ радиационной гигиены им. П.В. Рамзаева и СПбГУ проектов, посвященных исследованию различных социальных явлений, и выявление общности методов разработки, влияющих на эффективность решения аналитических задач при эксплуатации создаваемых АИС и БД.

Задачи исследования

1. Выявить свойства социальной информации и содержащих ее интернет-ресурсов, определяющие возможности и ограничения ее систематизации, классификации, поиска и взаимосвязи.

2. Определить требования к функциональным возможностям АИС и БД, необходимым для достижения поставленной цели при решении задач обработки и анализа данных.

3. Описать основные сложности при разработке АИС и БД, предназначенных для работы с социально-коммуникативными данными.

4. Сформулировать на основе накопленного опыта рекомендации по разработке АИС и БД для практического применения при решении задач в области коммуникации риска.

Материалы и методы

Анализ использованных при разработке АИС и БД подходов выполнен по методике автоэтнографического исследования [12]. Поскольку цель и задачи настоящего исследования носят преимущественно прикладной характер, возможности автоэтнографии применены прагматически – исключительно для анализа собственного опыта авторов по созданию и внедрению исследовательских АИС и БД для целей конкретных проектов.

Объект анализа – два случая разработки и использования АИС и БД, независимо реализованные исследовательскими коллективами НИИ радиационной гигиены им. П.В. Рамзаева и СПбГУ для решения задач систематизации и анализа публикаций. Рассмотрение двух независимых проектов в рамках одной статьи позволило выделить особенности формирования структуры БД информационных материалов, определяемые различной природой таких материалов. В проекте НИИ радиационной гигиены задача структурировать и классифицировать информационные материалы относится главным образом к электронным версиям традиционных СМИ. Проект СПбГУ ориентирован на более широкий круг документов, опубликованных в сети Интернет (включая комментарии к публикациям в СМИ, форумы, блоги и другие ресурсы). Применение стратегии сравнительного анализа позволяет как выявить общие черты, характеризующие ситуацию

разработки и внедрения АИС и БД в практику анализа информационного поля, так и описать различия в подходах, связанные с особенностями разных видов информационных материалов.

Первый случай – разработка опытной автоматизированной системы анализа публикаций по вопросам радиационной безопасности (РБ). Анализ публикаций осуществлялся в рамках работы по исследованию информационного поля в различных регионах Российской Федерации по вопросам радиационной безопасности. Работа по созданию АИС была начата в 2015 г. В настоящее время система содержит 1911 публикаций в 3 субъектах Российской Федерации за период с 1 октября 2016 г. по 31 марта 2017 г. для Мурманской области и с 1 января 2016 г. по 30 сентября 2016 г. для Санкт-Петербурга и Ленинградской области.

Второй из рассматриваемых случаев – коллекция размещенных в сети Интернет материалов по теме миграции. Работы по ее систематическому сбору начаты коллективом исследователей СПбГУ в 2010 г., когда для анализа общественного мнения Санкт-Петербурга о миграции и мигрантах было собрано 395 документов объемом от 700 до 1 800 000 знаков.

Результаты и обсуждение

Краткое описание АИС анализа публикаций по вопросам РБ

Автоматизированная информационная система по анализу публикаций (АСАП) была разработана специалистами НИИ радиационной гигиены им. П.В. Рамзаева для решения задач по анализу информационного поля регионов реализации мероприятий федеральной целевой программы «Обеспечение ядерной и радиационной безопасности на 2016–2020 годы и на период до 2030 года» [13]. АСАП была разработана в качестве подсистемы автоматизированной системы контроля радиационного воздействия Роспотребнадзора [14]. АСАП была создана на платформе «1С: Предприятие 8.3» (Россия), а в качестве СУБД был выбран Microsoft SQL Server 2012 (США). 1С: Предприятие обладает развитым аналитическим инструментарием, поддерживающим использование различных средств визуализации (графики, схемы, диаграммы, таблицы, географические карты и др.). В АСАП реализована возможность работы пользователей через web-браузер, то есть для работы в системе нет необходимости устанавливать на компьютер пользователя какое-то дополнительное программное обеспечение (ПО). Этим существенно повышается удобство работы, т.к. основным источником анализируемой информации являются web-сайты традиционных СМИ и информационных агентств, и у пользователя отсутствует необходимость переключения между программами при обнаружении необходимой информации. АСАП является многопользовательской системой и включает в себя две основных подсистемы – подсистему учёта публикаций и аналитическую подсистему. Подсистема учета публикаций содержит информацию о каждой найденной публикации. Помимо хранения текста публикации, возможности сохранения электронной копии, дополнительных материалов, ссылок на расположение материала в сети Интернет и данных, относящихся к идентификации материала (наименование СМИ, вид СМИ, дата публикации, наименование публикации и т.п.), система позволяет хранить результаты классификации материалов по специально разработанным для целей автоматического анализа классификаторам (тематика публикации (атомная энергетика, аварии, Чернобыль, медицина и т.д.); действующие лица публикации (население, ликвидаторы, чиновники, специалисты, общественные объединения); жанр публикации (информационный, аналитический, художественно-публицистический) с дополнительным уточнением (интервью, статья, колонка, заметка и т.д.); характер представления информации (нейтральный, негативный, позитивный); территория (Российская Федерация, субъект Российской Федерации, зарубежные страны)).

Разработка БД социальной информации в сети Интернет по вопросам миграции

Коллекция материалов, собранных социологами СПбГУ в 2010 г., включала в себя документы, размещенные (и/или активные в случае дискуссионных площадок) в двухнедельный период на сайтах, представляющих четыре основные категории ресурсов – информационные (СМИ), официальные (государственные и аффилированные с органами управления), специализированные по теме миграции (в том числе общественных организаций) и неспециализированные (форумы и блоги). Все собранные документы были интегрированы в общую базу, включавшую сами документы в едином текстовом формате MS Word с сохранением фотографий, рисунков, схем, присутствующих в тексте, а также сопроводительную информацию об источнике (включая ссылки по ГОСТ), времени сбора, авторстве (включая специально разработанную классификацию) и другие сведения, в том числе ссылки на другие документы и категории ресурсов. Созданная коллекция позволила успешно решить задачи вначале качественного (секвенционального), а затем количественного (по методике, предложенной Д.П. Гаврой [15]) анализа общественного мнения, представленного в русскоязычном сегменте сети Интернет.

Хотя описанная выше первая коллекция интернет-документов была собрана преимущественно вручную (исключение составило применение инструментов автоматического поиска), она послужила прототипом для разработки аналогичных баз, позволяющих увеличить объемы и сократить время сбора исследовательских материалов. В частности, для автоматизации сбора документов с ресурсов, характеризующихся высокой степенью стандартизации публикаций, эффективным признано применение программы Content Downloader (CD) (Россия) [16]. Так, в октябре 2017 г. с помощью CD была сформирована коллекция материалов по теме миграции, опубликованных на сайте RBC*, – всего 1696 документов за период 2002–2017 гг.

* РБК лента новостей. ЗАО «РОСБИЗНЕСКОНСАЛТИНГ». – Available on: https://www.rbc.ru/ (accessed: November 07, 2017) [RBC news feed. RosBusinessConsulting CJSC]

Анализ использованных подходов к разработке структуры данных

Для автоматизации обработки информационных материалов решающее значение имеет то, соответствует ли структура базы данных, предназначенная для хранения таких материалов и сведений о них, задачам будущего анализа. Рассмотрим особенности первичных данных, которые собираются в Интернете в рамках социальных и коммуникационных исследований. Прежде всего это текстовый, дискурсивный характер большинства материалов. В рассматриваемых нами случаях это публикации в СМИ, документы, размещенные на официальных сайтах и сайтах общественных организаций, выступления государственных, политических и общественных деятелей, продукты общественной интернет-дискуссии, результаты экспертных интервью и массовых опросов и т.д. Все вместе они образуют тематические коллекции материалов, с которыми предстоит работать исследователям. Причем чем масштабнее задачи анализа и шире охват источников, тем больше разнообразие обрабатываемых материалов. Так, анализ публикаций по теме радиационной безопасности проводился только по материалам электронных версий СМИ, чего было достаточно для решения задач исследования, тогда как для исследований публикаций о миграции собирались материалы из более широкого круга источников.

Коллекции разноплановых текстовых документов характеризуются рядом параметров, которые, с одной стороны, увеличивают объем базы данных и повышают требования к возможностям вычислительной системы, а с другой – наделяют материал рядом характерных черт, важных при проектировании баз данных. Одной из них является смешанный характер коллекции, с преобладанием текстовых массивов. При этом контент характеризуется сильным информационным шумом. К этому добавляется высокая зависимость от контекста.

Из сказанного вытекают два требования. Первое: необходимость сохранять информационные фрагменты большего объема, чем единицы анализа, так как отсечение части текста, создающей информационный шум, может частично или полностью исказить или уничтожить смысл единицы анализа. Второе: необходимо уже на этапе моделирования закладывать ресурсы для сохранения сопутствующих элементов кода или текста – фото-, аудио-, видеоматериалов, гиперссылочного аппарата материалов, так как все это контекстуально по отношению к единицам анализа. В случае с экспертными интервью возникает необходимость полностью или частично сохранять аудиозаписи, а в случае с материалами СМИ может потребоваться полностью или частично сохранять видеоряды, сопутствующие выявленным единицам анализа.

Сказанное выше в равной мере относится к подходам НИИ радиационной гигиены и СПбГУ. Задачи следующего этапа были решены по-разному.

При разделении задач анализа и хранения данных, как это было сделано в проекте СПбГУ, необходимо предусмотреть формирование многоуровневых баз данных, связанных между собой этапами преобразования материалов. Приведем пример с исследованием материалов СМИ. На первом этапе с сайтов популярных информационных изданий собираются материалы с упоминанием исследуемой тематики – с любым содержательным фокусом. На втором этапе эти материалы нужно классифицировать. Модифицировать первоначальный архив нерационально, так как он дает полный, одномоментный срез информационного пространства. Значит, надо сформировать базу данных следующего уровня, где к материалам добавляется новый атрибут – характер упоминания темы. На следующем этапе происходит более серьезное преобразование материалов, в ходе которого полные тексты новостных статей дефрагментируются на отдельные категории анализа. Делать это ни в первом, ни во втором архиве также нерационально, так как в любой момент может возникнуть необходимость обратиться к целому тексту или его фрагменту для уточнения смысловых единиц. Таким образом, даже в такой упрощенной схеме у нас получается три уровня данных, которые должны сохранять взаимосвязь между собой.

В АИС НИИ радиационной гигиены структура единицы анализа (публикации) такова, что включает в себя одновременно и исходные данные публикации (независимо от формата данных), и результаты классификации публикации по всем используемым классификаторам. Особенность подхода заключается в том, что добавление новых параметров классификации не затрагивает уже введенные в систему данные, а реальная структура БД остается «внутренним» вопросом среды разработки. То есть разработчик сосредоточен прежде всего на практической стороне вопроса, а не на технических аспектах реализации БД.

Подходы к решению аналитических задач, как следует из вышесказанного, тоже отличаются. БД СПбГУ ориентирована на использование внешних программ для анализа введенных данных. Преимуществом такого подхода является свобода выбора внешней программы с возможностью использования той, чей функционал наиболее пригоден для решения конкретной аналитической задачи. Слабым местом такого подхода являются более сложные технические требования к структуре разрабатываемой БД. Количественный анализ при таком подходе требует создания на основе уже имеющихся в коллекции материалов принципиально иного по своей форме продукта – электронной таблицы, с которой может работать предпочитаемая аналитиком статистическая программа – SPSS (США), SAS (США), Statistica (США) и др. Здесь, во-первых, обратим внимание на ограничения «конечной» программы обработки данных, т.к. даже простой перенос данных из таблицы одного формата в таблицы другого формата на практике нередко реализуем только в ручном режиме и требует внимательной перепроверки, корректировки, восстановления потерянных данных и т.п. «Идеальная» же исследовательская БД должна обеспечивать аналитика работоспособным массивом данных в требуемом формате автоматически, а во многих случаях – еще и обеспечивать автоматическое обновление. Этого недостатка лишен подход НИИ радиационной гигиены, т.к. изначально предполагается использование встроенных средств анализа данных. Слабым местом подхода является то, что возможности встроенного аналитического инструментария 1С: Предприятия уступают таковым у специализированных пакетов статистической обработки данных. При этом следует отметить, что не существует ограничений по экспорту данных во внешние

электронные таблицы любого формата для использования внешними программами. В этом случае, однако, перед разработчиками АИС встанут ровно такие же задачи обеспечения совместимости форматов данных.

При разделении задач анализа и хранения данных между коллекцией текстовых материалов и электронной таблицей стоит еще одна процедура – кодировка. В «классических» социальных исследованиях кодировка выполнялась вручную, что с известной мерой условности (связанной с уровнем подготовки, мастерства, мотивации исследователей) обеспечивало ее эффективность – человек всегда может принять решение о том, какой код присвоить очередной единице анализа. Но ручная кодировка губительная для полноценных БД – теряется значительная часть их преимуществ. Кроме того, это очень ресурсоемко. Например, при обработке первой коллекции по миграционной тематике материалы для кодировки были сгруппированы в 17 файлов – каждый объемом около 350 000 знаков. Работы велись группой из 14 кодировщиков в течение 2 недель (неполный рабочий день), что позволило закодировать 2009 единиц – самостоятельных высказываний, содержащих мнения по предмету исследования.

Автоматизация процедур кодирования текстовых данных – необходимая мера при разработке БД. На данный момент, несмотря на наличие различных программных продуктов (ATLAS.ti (Германия), QDA Miner (Канада), Concordance (Великобритания)) [17], эффективно решаются только относительно простые задачи кодирования данных. Одной из главных проблем является многообразие дискурсивных практик, в том числе сленговых, субкультурных и контекстуальных. Исследователям приходится отслеживать, какие вербальные конструкции используются авторами суждений, своевременно дополняя используемые списки ключевых слов.

Другая ответственная задача носит технический характер: разработка данного компонента БД предполагает необходимость согласовать формат, в котором сохранены собранные дискурсивные данные с форматами данных программного обеспечения, используемого для кодирования и последующей статистической обработки. Нетрудно догадаться, что ни журналист, ни блогер, ни «форумчанин» не заботятся о том, чтобы их тексты были удобны для контент-анализа. Существует и другое мнение по этому поводу, состоящее в том, что для задач контент-анализа нельзя создать (или использовать) унифицированную программу обработки. Программы предлагают широкий функционал, а задачи анализа всегда практические и конкретные. Поэтому приложение нужно писать самостоятельно, то есть исследователь должен владеть навыками программирования или включать программиста в исследовательскую команду.

И еще один важный момент – количественный анализ эффективен, когда применяются методически обоснованные аналитические инструменты, позволяющие подняться над распространенным, к сожалению, в современной социологии анализом отдельных распределений. Например, в исследованиях общественного мнения о миграции коллектив СПбГУ успешно использует систему показателей, разработанную петербургским социологом Д.П. Гаврой. Но это означает, что БД должна обеспечивать не просто кодировку определенных переменных, пригодных для статистической обработки, но и поддерживать включение этих переменных в аналитические расчеты, что, естественно, накладывает дополнительные ограничения и на данные, и на процедуры.

Оба рассматриваемых подхода предполагают, помимо внимания к контексту, и учет того, что сами материалы находятся внутри информационного поля, в котором они созданы. Ресурсы, авторы, участники и комментаторы создают сегменты в информационном пространстве. Кто-то за, кто-то против. Кто-то высказывается по данному вопросу один раз, кто-то, как определенные издания или журналисты, могут заниматься этой темой и создавать информационные артефакты годами, внося тем самым значимые изменения в контекст исследования. Это значит, что кроме внутренней взаимосвязи материалов на разном уровне обработки, они должны сохранять взаимосвязь с другими полями (материалами, источниками, авторами), одним словом – все те взаимосвязи, которые удалось обнаружить в ходе исследования. Таким образом получаются многомерные, иерархические, текстовые базы данных, с которыми работает коллектив авторов, каждый из которых, изучая свой тематический аспект, работает с полной базой данных. Это значит, что при проектировании баз данных нужно понимать не только характер данных, но и операции с данными, так как каждый новый параметр может потребовать внесения изменений в БД [18].

К числу подобных операций относится сравнение материалов. Это связано с тем, что в информационном пространстве тексты кочуют из источника в источник. Разные информационные издания могут перепечатывать не только одну новость, но и один текст, а участники дискуссии – перепечатывать высказывания друг друга или печатать одно и то же в разных дискуссиях, например на разных форумах. Сравнение текстов внутри базы данных позволит сгруппировать такие источники и выявить «кочующие» в информационном пространстве тексты. Это полезно еще и потому, что тогда можно анализировать, какую реакцию подобные повторы вызывают в разных аудиториях. Вторая задача – убирать дублирующиеся материалы из базы данных. Когда собираются коллекции, насчитывающие десятки и сотни материалов, повторы неизбежны.

Еще одна особенность информационного поля, которую необходимо учитывать при определении структуры данных, – это его изменчивость. Авторы могут редактировать свои публикации, информационные источники, новостные статьи, участники форумов – свои высказывания, модераторы сайтов могут удалять как отдельные фразы, так и целые материалы или дискуссии. Все это сказывается на контексте единиц анализа и на восприятии материалов аудиторией. Таким образом, сравнение необходимо не только на предмет совпадения материалов между собой в момент ввода данных, но и между массивами, сформированными в ходе повторных этапов поиска и архивирования информации. Это позволяет увидеть, как именно материалы меняются во времени, и разобраться, с чем эти изменения связаны, на чем лежит локус контроля в общественном информационном пространстве. Еще одно следствие изменчивости информационного пространства – могут безвозвратно исчезать и уничтожаться целые сайты, и своевременное архивирование информации позволит не потерять важные и ценные данные [19].

При проектировании базы данных и оптимизации СУБД под исследовательскую задачу существует опас-

ность впасть в одну из крайностей. В первом варианте можно формализовать все поля базы данных, сохраняя только интересующее нас информационное ядро (единицу анализа) и при этом полностью избавиться от информационных шумов и контекстного сопровождения данных, тем самым упростив до минимума все взаимосвязи между данными, так как они будут неочевидны. Таким образом можно получить легкую в отношении ресурсов хранения, простую в управлении и удобную базу. Она будет наполнена очень простыми, практически примитивными смыслами, но довольно тривиальным содержанием, и в значительной степени лишена исследовательского смысла. Другой вариант, напротив, предполагает попытку сохранения всего информационного среза полностью, со всеми смысловыми, оттеночными нюансами контекста и взаимосвязями между единицами анализа. В этом случае получится очень большая, «неповоротливая» база данных с крайне сложной структурой. Такой подход к структурированию информации и сопровождающих ее шумов не позволит вычленить значимые элементы, и единицы анализа растворятся в общем массиве информации.

Оптимальная архитектура базы данных исследовательского проекта — это сложный компромисс между информационной полнотой содержания данных и технической возможностью их обработать. При работе с таким типом материалов, которые были описаны выше, не все операции можно эффективно автоматизировать. Интеграция интеллектуальных систем анализа позволит автоматизировать только самые простые операции с данными. А интернет-пространство – неструктурированное и поливариантное. Кроме изменчивости, материалы наследуют от него различную структуру представления информации.

Материалы с информационного сайта, с форума и собранные в публичных дискуссиях социальных сетей не только не совпадают по структуре между собой, они еще и различным образом систематизированы в структуре сайта. Разные новостные сайты не только поддерживают разные позиции и установки – они имеют собственную структуру и в подаче материала, и в верстке сайта. А на одном и том же форуме (или в сообществе/группе социальной сети) по одному и тому же вопросу можно найти дискуссии (ветки форумов/обсуждения) разной длины. Полностью автоматизировать обработку таких разных по целому ряду параметров данных не получится, по крайней мере в настоящее время и в ближайшей перспективе развития программного обеспечения. Следовательно, глубокий «ручной» анализ первичных исходных данных является важнейшей задачей, предшествующей разработке АИС и БД, и в значительной степени определяет эффективность создаваемого инструмента анализа.

Выводы

1. Сравнительный анализ результатов работы двух независимых друг от друга научных коллективов по разработке и внедрению АИС и БД, предназначенных для решения исследовательских задач анализа информационных материалов в области риск-коммуникации, показывает, что методические подходы имеют много общего. В частности, наблюдается общность подходов при решении задач классификации и систематизации материалов. В значительной степени это определяется спецификой социальной информации, представленной в Интернете – преобладанием текстовых документов, в которых контекст имеет большое значение для интерпретации информации.

2. Материалы, относящиеся к описанию социально значимых явлений, отличаются информационной насыщенностью, разнообразием ресурсов и форматов представления документов. Поэтому уже на этапе разработки архитектуры АИС и структуры БД необходимо ориентироваться на создание достаточно больших массивов разнородной, но взаимосвязанной текстовой, графической и аудиовизуальной информации, структурированной в пригодном для последующего анализа виде. Помимо хранения больших массивов, требуется организовать в них внутренние связи и обеспечить возможность последовательного создания новых уровней БД, предусматривающих решение задач анализа информации.

3. Важным условием при создании эффективных, с точки зрения социально-коммуникационных исследований, АИС и БД является реализация возможности обработки собранных социальных данных с помощью доступных исследователям программ качественного и количественного анализа, если такой анализ не может быть осуществлен собственными средствами АИС.

4. База данных информационных материалов – это компромисс между полнотекстовой коллекцией и аналитическими возможностями системы, а также между оптимизацией под задачи исследования и унификацией под требования аналитических систем. Требование автоматизации ряда функций упирается в высокую степень уникальности материалов, которая пока де-факто преодолевается только ручной пре- и постобработкой, что необходимо учитывать при проектировании исследовательских программ.

Статья подготовлена в ходе реализации проекта «Создание модели многофункционального центра компетенций в области социальной работы с мигрантами в условиях нарастания потока переселенцев в Россию и Швейцарию для снижения угроз обществу, экономике, государству», реализуемого в Санкт-Петербургском государственном университете при поддержке Федеральной целевой программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014–2020 годы», уникальный идентификатор проекта RFMEFI61317X0072.

Автоматизированная информационная система по анализу публикаций (АСАП) была разработана в ходе выполнения работ по государственному контракту № Н.4Д.241.20.17.1026 от 20 марта 2017 г. по теме «Разработка и научное обоснование практических мероприятий по освещению в Северо-Западном федеральном округе Российской Федерации деятельности по повышению радиационной безопасности в рамках федеральной целевой программы «Обеспечение ядерной и радиационной безопасности на 2016–2020 годы и на период до 2030 года» в обеспечение мероприятия «Разработка методических основ и организация информационной работы с населением по вопросам радиационной безопасности в районах реализации мероприятий Программы»» в рамках федеральной целевой программы «Обеспечение ядерной и радиационной безопасности на 2016–2020 годы и на период до 2030 года».

Литература

1. Архангельская, Г.В. Проблемы риск-коммуникации по вопросам радиационной безопасности: оценка информированности населения Санкт-Петербурга и Ленинградской области о деятельности атомной отрасли и его представления о факторах опасности / Г.В. Архангельская, С.А. Зеленцова, Н.М. Вишнякова, Е.В. Храмцов, К.В. Варфоломеева, Н.В. Соколов, В.С. Репин // Радиационная гигиена. – 2017. – Т. 10, № 3. – С. 36–45.
2. Соколов, Н.В. Проблемы риск-коммуникации при обеспечении радиационной безопасности: представление о радиации и атомной отрасли в массовом сознании по результатам социологических исследований в Санкт-Петербурге, Ленинградской и Мурманской областях / Н.В. Соколов, А.М. Библин, Л.В. Репин, Л.С. Рехтина // Радиационная гигиена. – 2017. – Т. 10, № 3. – С. 45–56.
3. Углев, С.В. Обзор систем синтаксического анализа и отладки хранимых процедур в различных СУБД / С.В. Углев // Вестник МГУП. – 2016. – № 2. – С. 69–71.
4. Мишанкина, Н.А. Базы данных в лингвистических исследованиях / Н.А. Мишанкина // Вопросы лексикографии. – 2013. – №1 (3). – С. 25–33.
5. Espinoza M.D., Datcu M. Data mining and knowledge discovery tools for exploiting big earth observation data International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences – ISPRS ArchivesVolume 40, Issue 7W3, 28 April 2015, Pages 627-633.
6. Bertke S.J., Meyers A.R., Wurzelbacher S.J., Measure A., Lampl M.P., Robins D. Comparison of methods for auto-coding causation of injury narratives Accident Analysis and PreventionVolume 88, 1 March 2016, Pages 117-123.
7. *Бек, У.* Общество риска. На пути к другому модерну / *У. Бек*. – М.: Изд-во: Прогресс-Традиция, 2000. – 480 с.
8. Репин, В.С. Дозы облучения населения Российской Федерации по итогам функционирования ЕСКИД в 2002–2015 гг. : информ. сборник / В.С. Репин, Н.К. Барышков, А.А. Братилова [и др.]. – СПб., 2015. – 40 с.
9. Соколов, Н.В. Работящие, но чужие: парадокс восприятия мигрантов массовым сознанием (по результатам исследований в Санкт-Петербурге) / Н.В. Соколов // Мониторинг общественного мнения: экономические и социальные перемены. – 2017. – № 1 (137). – С. 80–96.
10. Эриксен, Т.Х. Тирания момента. Время в эпоху информации / Т.Х. Эриксен; пер. с норв. – М.: Весь Мир, 2003. – 208 с.
11. Шаркова, Е.А. Коммуникация в условиях экологического риска / Е.А. Шаркова // Вестник СПбГУ. Серия 9. Филология. Востоковедение. Журналистика. – 2011. – №4. – С. 237–245.
12. Рогозин, Д. Как работает автоэтнография? / Д. Рогозин // Социологическое обозрение. – 2015. – Т.14, № 1. – С. 224–273.
13. Библин, А.М. Анализ характера освещения в средствах массовой информации радиационной безопасности населения Санкт-Петербурга и Ленинградской области / А.М. Библин // Радиационная гигиена. – 2017. – Т. 10, № 2. – С. 23–30.
14. Репин, Л.В. Автоматизированная система контроля радиационного воздействия Роспотребнадзора: история создания, назначение и развитие / Л.В. Репин, А.М. Библин, П.Г. Ковалев, М.С. Николаевич, В.С. Репин // Радиационная гигиена. – 2015. – Т. 7, № 3. – С. 44–53.
15. Гаврa, Д.П. Общественное мнение как социологическая категория и социальный институт / Д.П. Гаврa. – СПб.: ИСЭП РАН, 1995.
16. Content Downloader X1 версии 11.1.0000233 (02.09.2017): http://sbfactory.ru (дата обращения: 07.11.2017).
17. Рюмин, А. Блог о контент-анализе. Софт / А. Рюмин. – http://content-analysis.ru/index.php/luchshij-soft-dlya-kontent-analiza (дата обращения: 09.11.2017).
18. Мокрозуб, В.Г. Синтаксис запросов конечных пользователей к реляционной базе данных / В.Г. Мокрозуб // Прикладная информатика. – 2009. – № 3. – С. 95–99.
19. Бегтин, И. Архивы государственных сайтов / И. Бегтин. – http://ivan.begtin.name/2012/03/23/govarchiv/ (дата обращения: 09.11.2017).

Поступила: 10.11.2017 г.

Рехтина Лилия Сергеевна – магистр социологии, сотрудник проекта, Санкт-Петербургский государственный университет, Правительство Российской Федерации, Санкт-Петербург, Россия. **Адрес для переписки:** 191124, Россия, Санкт-Петербург, ул. Смольного, д. 1/3, 9-й подъезд; E-mail: lisabet-09@mail.ru

Соколов Николай Викторович – кандидат социологических наук, доцент, Санкт-Петербургский государственный университет, Правительство Российской Федерации.

Библин Артем Михайлович – и.о. руководителя Информационно-аналитического центра, Санкт-Петербургский научно-исследовательский институт радиационной гигиены имени профессора П.В. Рамзаева Федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека, Санкт-Петербург, Россия.

Репин Леонид Викторович – младший научный сотрудник Информационно-аналитического центра Санкт-Петербургского научно-исследовательского института радиационной гигиены имени профессора П.В. Рамзаева Федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека, Санкт-Петербург, Россия

Ахматдинов Рустам Расимович – ведущий инженер-исследователь Информационно-аналитического центра Санкт-Петербургского научно-исследовательского института радиационной гигиены имени профессора П.В. Рамзаева Федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека, Санкт-Петербург, Россия

Для цитирования: Рехтина Л.С., Соколов Н.В., Библин А.М., Репин Л.В., Ахматдинов Р.Р. Проблемы аналитического обеспечения коммуникации рисков: обоснование подходов к разработке исследовательских баз данных по вопросам радиационной безопасности и социальных рисков // Радиационная гигиена. – 2017. – Т. 10, № 4. – С. 44-52. DOI: 10.21514/1998-426X-2017-10-4-44-52.

Analytical issues of risk communication. Rationale for approaches to developing research databases on radiation safety and social risks

Liliya S. Rekhtina[1], Nikolay V. Sokolov[1], Artem M. Biblin[2], Leonid V. Repin[2], Rustam R. Akhmatdinov[2]

[1]Saint-Petersburg State University, The Government of the Russian Federation Saint-Petersburg, Russia

[2]Saint-Petersburg Research Institute of Radiation Hygiene after Professor P.V. Ramzaev, Federal Service for Surveillance on Consumer Rights and Human Well-Being, Saint-Petersburg, Russia

One of the important stages of risk communication is the analysis of publications in traditional media and the Internet, which largely shape people's attitudes to various issues. At the same time, the availability of large amounts of information relating to any subject area complicates the possibility of manual analysis and adequate description of all of the information. On the other hand, the availability of information causes the urgency of developing methods to improve the effectiveness of its analysis. One way to automate the analysis of large amounts of information is the development of databases or automated information systems containing information materials on the subject matter under study and suggesting the possibility of automated processing. The objective of this work is to analyze the experience of developing such systems and databases by the research teams of the St. Petersburg Institute of Radiation Hygiene and St. Petersburg State University and to identify key features of the use of bases Data for social research. The results of the analysis showed that the methodological approaches used were very close. The analysis is performed according to the method of autoethnographical research. The strategy application of the comparative analysis allows identifying common features characterizing the situation of development and implementation of databases to practice of the risk communication studies. The article discusses the features associated with them, the limitations of the primary data, such as text, discursive nature of most of the materials, information noise, high dependence on context, variability, different structure, format and appearance of materials. The important parameters for solving problems of the qualitative and quantitative analysis are given in the article. An important condition of creating effective, from the point of view of socio-communication studies information system is to implement the processing capabilities of the collected social data available to researchers programs. The requirement for automation of certain functions depends on a high degree of uniqueness of the materials, which is overcome only by the manual pre- and post-processing, which must be considered when designing research programs.

Key words: *databases, automated information systems, radiation safety, risk communication, analysis of publications, social risks, sociology of science, sociology of migration.*

References

1. Arkhangelskaya G.V., Zelentsova S.A., Vishnyakova N.M., Khramtsov E.V., Varfolomeeva K.V., Sokolov N.V., Repin V.S. Risk-communication issues in radiation safety: evaluation of public awareness in St. Petersburg and the Leningrad region on the activities of the nuclear industry and public understanding of the hazards. Radiatsionnaya gygiena = Radiation Hygiene, 2017, Vol. 10, № 3, pp. 36-45. (In Russian).
2. Sokolov N.V., Biblin A.M., Repin L.V., Rekhtina L.S. Risk-communication issues in radiation safety: Mass consciousness about radiation and nuclear industry based on the results of a sociological research in St. Petersburg, the Leningrad region and the Murmansk region. Radiatsionnaya gygiena = Radiation Hygiene, 2017, Vol. 10, № 3, pp. 45- 56. (In Russian).
3. Uglev S.V. Review of systems for parsing and stored procedure debugging in different DBMS. Magazine Vestnik MGUP, 2016, №2, pp. 69-71. (In Russian).
4. Mishankina N.A. Databases in linguistic research. Voprosy leksikografii = Russian Journal of Lexicography, 2013, №1 (3), pp. 25-33. (In Russian).
5. Espinoza M.D., Datcu M. Data mining and knowledge discovery tools for exploiting big earth observation data International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences – ISPRS ArchivesVolume 40, Issue 7W3, 28 April 2015, Pages 627-633.
6. Bertke S.J., Meyers A.R., Wurzelbacher S.J., Measure A., Lampl M.P., Robins D. Comparison of methods for auto-coding causation of injury narratives Accident Analysis and PreventionVolume 88, 1 March 2016, Pages 117-123.
7. Bek U. Risk Society: Towards a New Modernity. Moscow, Progress-Tradicija, 2000, 384 p. (In Russian).
8. Repin V.S., Baryshkov N.K., Bratilova A.A. [et al.]. Information packet: Radiation exposure doses of the population of the Russian Federation according to the results of the USIDC in 2002-2015. SaintPetersburg, 2015, 40 p. (In Russian).
9. Sokolov N.V. Hard-working but outsiders: paradox of perception of migrants in mass consciousness (the case of St. Petersburg). The Monitoring of Public Opinion: Economic and Social Changes Journal, 2017, № 1 (137), pp. 80-96. (In Russian).
10. Eriksen T.Kh. Tyranny of the Moment: Fast and Slow Time in the Information Age. Moscow, 2003, 208p. (In Russian).

Liliya S. Rekhtina
Saint-Petersburg State University.
Address for correspondence: Smolnogo str., 1/3, 9th entrance, St. Petersburg, 191124, Russia; E-mail: lisabet-09@mail.ru

11. Sharkova E.A. Communication in the environmental risk context. Vestnik St.Petersburg University, Ser. 9, 2011, Issue 4, pp. 237–245. (In Russian).

12. Rogozin D. How Autoethnography Works. Sotsiologicheskoe obozrenie = Russian Sociological Review, 2015, Vol. 14, № 1, pp.224-273. (In Russian).

13. Biblin A.M. Analysis of the media coverage characteristics on radiation safety issues of the Saint-Petersburg and the Leningrad region population. Radiatsionnaya gygiena = Radiation Hygiene, 2017, Vol. 10, №. 2, pp. 23-30. (In Russian).

14. Repin L.V., Biblin A.M., Kovalev P.G., Nikolaevich M.S., Repin V.S. The automated system of radiation exposure control for Rospotrebnadzor: creation history, applicability and development. Radiatsionnaya gygiena = Radiation Hygiene, 2014, Vol. 7, No 3, pp. 44-53. (In Russian).

15. Gavra D.P. Public opinion as a sociological category and social institution. Saint-Petersburg, ISJeP RAN, 1995. (In Russian).

16. Content Downloader X1 version 11.1.0000233 (02.09.2017). -Available on: http://sbfactory.ru (accessed: November 07, 2017). (In Russian).

17. Ryumin A. Blog about content analysis. Software. -Available on: http://content-analysis.ru/index.php/luchshij-soft-dlya-kontent-analiza (accessed: November 09, 2017). (In Russian).

18. Mokrozub V.G. The syntax of end user requests to a relational database. Applied informatics, 2009, №3, pp. 95-99. (In Russian).

19. Begtin I. Archives of state websites. -Available on: http://ivan.begtin.name/2012/03/23/govarchiv/ (accessed: November 09, 2017). (In Russian).

Received: November 10, 2017

For correspondence: Liliya S. Rekhtina – Master of Sociology, Project Staff, Saint-Petersburg State University (Smolnogo str., 1/3, 9th entrance, St. Petersburg, 191124, Russia; E-mail: lisabet-09@mail.ru)

Nikolay V. Sokolov – Candidate of Sociological Science, Assistant Professor, Saint-Petersburg State University, Saint-Petersburg, Russia

Artem M. Biblin – Information and Analytical Center Head, Saint-Petersburg Research Institute of Radiation Hygiene after Professor P.V. Ramzaev, Federal Service for Surveillance on Consumer Rights Protection and Human Well-Being, Saint-Petersburg, Russia

Leonid V. Repin – Juonior Researcher, Information and Analytical Center, Saint-Petersburg Research Institute of Radiation Hygiene after Professor P.V. Ramzaev, Federal Service for Surveillance on Consumer Rights Protection and Human Well-Being, Saint-Petersburg, Russia

Rustam R. Akhmatdinov – Leading Engineer Researcher, Information and Analytical Center, Saint-Petersburg Research Institute of Radiation Hygiene after Professor P.V. Ramzaev, Federal Service for Surveillance on Consumer Rights Protection and Human Well-Being, Saint-Petersburg, Russia

For citation: Rekhtina L.S., Sokolov N.V., Biblin A.M., Repin L.V., Akhmatdinov R.R. Analytical issues of risk communication. Rationale for approaches to developing research databases on radiation safety and social risks. Radiatsionnaya gygiena = Radiation Hygiene, 2017, Vol. 10, No 4, pp. 44-52. (In Russian) DOI: 10.21514/1998-426X-2017-10-4-44-52.