

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ
ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г.

№ 9

Москва 2016

ОБЩИЙ РАЗДЕЛ

УДК 316.77 : [001.102 : 002.1] – 027.21

Е.А. Плешкевич

Концепция социальных коммуникаций Никласа Лумана и теория документальной информации: точки пересечения

Рассматривается концепция социальных коммуникаций Никласа Лумана. Показана взаимосвязь коммуникационных представлений с теорией информации. Проанализирован эвристический потенциал теории Н. Лумана применительно к дисциплинам документально-информационного цикла.

Ключевые слова: Никлас Луман, теория социальных коммуникаций, коммуникация, теория информации, библиотековедение, документально-информационный социальный институт

ВВЕДЕНИЕ

Поиск оснований современного общества – одна из ключевых задач социологии. В условиях определения общества как информационного неминуемо возникает вопрос о связи его оснований с информацией и коммуникацией. Следует заметить, что истоки этой связи можно проследить еще в работах Аристотеля, предлагавшего рассматривать человека как

«политическое животное, одаренное речью». Однако научный анализ связей общества с информационными и коммуникационными процессами становится возможным лишь в ходе разработки соответствующих теоретических конструкций. Одной из первых таких конструкций стала математическая теория связи Шеннона–Уивера и кибернетика Н. Винера, где были сформированы первые научные представления об информации и коммуникации. Спустя несколько

десятилетий эти понятия вошли в социологию. На их основе, в сочетании с системным подходом, немецким социологом Никласом Луманом (1927-1998) [1, 2] была разработана новая концепция общества. Если прежде общество рассматривалось как совокупность территориально объединённых индивидов и их отношений, то Н. Луман предложил посмотреть на него как на всеобъемлющую систему коммуникаций. Эти исследования получили высокую оценку научного мира. В [3] было отмечено, что новая концепция организации общества позволяет лучше описать феномен социальных систем всех уровней, по сравнению с предложенной Т. Парсонсом теорией социального действия, что его система понятий с той или иной степенью убедительности способна описать общество с помощью технологического языка информационной революции [4]. Признание научной значимости теории Лумана способствовало росту интереса к ней не только социологов, но и представителей других наук [5-8], а также ученых, занимающихся документально-информационной проблематикой и оперирующих информационными и коммуникационными понятиями, что, в свою очередь, и определяет актуальность выбранной нами темы.

КОММУНИКАЦИОННАЯ КОНЦЕПЦИЯ ОБЩЕСТВА Н. ЛУМАНА

В основе представлений Лумана об обществе лежит концепция самовоспроизводящихся систем, предложенная чилийскими нейрофизиологами У. Матурана и Ф. Варела. Под такими системами понимают системы, которые состоят из компонентов и отношений, способных воспроизводить и элементы, и связи между ними с помощью лишь собственных действий, т. е. операций, относящихся к самой системе, а не к окружающему ее миру. Самовоспроизводство или самореферентность оказывается возможным только при различии между системой и средой. При этом системой может быть все то, что способно к «автопоэзису», т. е. является самодостаточным и способно воспроизводить себя само, отличая себя от окружения. Социальные системы – это наиболее комплексные (сложные) системы из существующих, воспроизводящих себя с помощью коммуникации, направленной на обретение смысла как различия между системой и окружением. Это воспроизводство, трактуемое как автопоэзис, заключается в создании смысловых структур. В основе воспроизводства лежат коммуникационные процессы. Таким образом, можно сказать, что Луман апеллирует к смысловой коммуникации, к смыслу, имеющему три фундаментальных измерения: вещественное, временное и социальное. Вещественное измерение имеет различие между «тем и иным», что позволяет идентифицировать реальность как вещь. Временное измерение различает между «прошлым и будущим» и позволяет рассматривать реальность как временной процесс. Наконец, социальное измерение имеет различие между «Эго и Альтер эго». Оно позволяет дать всем проявлениям смысла адресные координаты, необходимые для существования социальных систем. В итоге общество рассматривается как система, конструируемая смыслом посредством коммуникации. Это радикальная смена взглядов: если

прежде общество виделось социологам как собрание индивидов, то сейчас оно видится как система, состоящая только из коммуникационных операций. Продуцирование социальной системой саму себя – это и есть лумановская коммуникация [4, с.164], которая при этом является процессом тройной селекции: информации, сообщения и понимания. Информация, трактуется Луманом как событие, которое определяет состояние системы [2, с.107]. Во-первых, проявляясь, она не исчезает бесследно, а оставляет структурный эффект, изменяя состояние системы. Во-вторых, информация не приводит к извне и не «передается», она может быть только результатом собственных операций самореферентальных систем, которые реагируют на изменение собственных состояний. В-третьих, информация переживается, а не добывается. Сообщение – это основной динамический элемент коммуникации, обуславливающий временное протекание коммуникации. В данном контексте сообщение – это селекция того, что должно получить распространение. И, наконец, в-четвертых, – понимание, которое трактуется как проведение различия между информацией и сообщением. В итоге понимание – это селекция из селекции: сначала мы выбираем что сказать, формируя сообщение, а после, из услышанного нами сообщения, выбираем информацию. «Как ваше здоровье, Абрам Израилевич? Не дождетесь!» – яркий пример такой коммуникативной ситуации.

Обобщая представления Лумана о коммуникации, можно отметить, что он дает ее расширенное понимание, которое мы предлагаем сопоставить со структурой информационного процесса [9]. Итак, коммуникация включает, во-первых, передачу некоего смыслового содержания от одного индивида к другому или от передатчика к приемнику, что можно рассматривать как синтаксический аспект коммуникации; во-вторых, «некое исторически-конкретное протекающее, зависимое от контекста событие». Очевидно, что событие может быть выделено, потому что оно имеет собственное, отличное от других содержание, это можно рассматривать как семантический аспект; в-третьих, коммуникация – это операция, в процессе которой происходит перераспределение знания и незнания, что можно квалифицировать как её прагматический аспект. В итоге, представления Лумана о коммуникации включают синтаксический, семантический и прагматический аспекты и таким образом фактически совпадают с представлениями об информационном процессе с той лишь разницей, что лумановский процесс протекает не в системах связи Шеннона, а в социальных структурах. А отсюда и акцент на семантическом и прагматическом аспектах, что позволяет лумановскую теорию коммуникации рассматривать как семантико-прагматическую.

Отталкиваясь от этих представлений о коммуникации, Луман формулирует идею медиакоммуникации, выделяя в ней средства распространения коммуникации и «символически генерализованные» средства коммуникации. К средствам распространения Луман относит письменность, книгопечатание, электронные средства коммуникации и обработки информации СМИ и, в определенной степени, рели-

гию и мораль. Средства распространения повышают достижимость адресатов, вероятность, что адресаты получают информацию, но они никак не гарантируют, что информация будет понята и принята, т. е. не гарантируют успех информации. С ростом переизбытка информации эта вероятность только уменьшается.

Завершенность, а, следовательно, и успешность коммуникации обеспечивается «символически генерализованными» средствами. К ним Луман относит истину, власть, право, любовь, искусство, ценности и т.д. Эти медиа благодаря своему бинарному «да-нет» коду, позволяют охватить всю реальность, относима к истине или лжи, к праву или несправедливости. Иными словами, в этой плоскости лежит понимание медиакоммуникации как завершающего элемента социальной коммуникации. Именно этот тип медиа связан с общественными функциональными системами: истина – с наукой, власть – с политикой, право – с юриспруденцией и т.д.

Каким же образом коммуникация формирует социальный порядок? Луман объясняет это через понятие «двойной контингентности». Термин «контингентность» буквально обозначающий случайность, Луман трактует как «не необходимость» и многофакторность внутрисистемных отношений. В самом простом смысле под «двойной контингентностью» следует понимать ситуацию, когда *Эго* и *Альтер* ориентируют свои действия на взаимные поведенческие ожидания. Иными словами – это «ожидание ожиданий». В этой ситуации, люди вынуждены вступать в коммуникации, уменьшая неопределенность и тем самым формируя социальный порядок. Именно поэтому рост контингентности сопровождается ростом значения порядка в системе. Несколько упрощая ситуацию, можно сказать, что рост неопределенности в имущественных отношениях привел к появлению права, в познании окружающего мира – к появлению религии и науки, потребность в воспроизведении потенциальных версий действительности – к появлению искусства и т.д.

ДОКУМЕНТАЛЬНО-ИНФОРМАЦИОННАЯ ИНТЕРПРЕТАЦИЯ ИДЕЙ ЛУМАНА

Что же нового вносит концепция Лумана в эти представления? Очевидно, что в рамках одной статьи раскрыть этот вопрос в полной мере не представляется возможным, в силу чего мы хотим ограничиться анализом отдельных ключевых положений.

Прежде обозначим основные трактовки понятия «коммуникации» в документально-информационных дисциплинах. Напомним, что в математической теории связи, предложенной К. Шенноном и У. Уивером, «коммуникация» трактовалась как передача сообщения от источника информации к адресату, а информация – как идеальная часть передаваемого сообщения, уменьшающая энтропию и снижающая неопределенность. В этом контексте они были введены в информатику, а затем в библиотечно-библиографическую науку. По ходу развития акцент был сделан на понятие коммуникации, которое получило свою интерпретацию в следующих аспектах. Во-первых, как социальная коммуникация, под которой предлагалось понимать движение смыслов в социальном времени и

пространстве [10, с. 22]; как способ (форма, канал) обмена смыслами в обществе, т.е. знанием, представляющим общественное или личностное значение [11, с. 331]. На основе этой интерпретации были предложены теория и метатеория социальных коммуникаций. Во-вторых, как документальные коммуникации – коммуникации, опосредованные документами [12, с. 15]. В документоведении такие коммуникации определяются как документные¹. И, в-третьих, как библиографическая коммуникация, заключающаяся во взаимодействии людей с помощью библиографических сообщений, – процесс обмена информацией о текстах, а также межсубъектная связь во время передачи библиографического знания. В результате этого взаимодействия происходит изменение ранее существовавшей вероятности выбора текстов у его субъектов [14, с.64]. Таким образом, в синтаксическом аспекте коммуникация трактуется преимущественно как распространение информации в форме смыслов, библиографических сообщений, а также в прагматическом аспекте – как то, что уменьшает неопределенность.

Что же нового в эти представления вносят идеи Лумана? Начнем с анализа связи технологий коммуникации и социального «конструирования». Как известно, развитие социума происходит как по вертикали, так и по горизонтали. Вертикальная дифференциация носит иерархический характер и направлена на выстраивание управления, тогда как горизонтальная дифференциация направлена на выделение и развитие отдельных сфер деятельности (производство, культура, экономика и т.д.). В связи с этим выскажем предположение, что вертикальное социальное конструирование шло по пути разработки новых «прорывных» коммуникационных технологий, обеспечивающих усиление властной вертикали. В Египте было изобретено сакрализованное иероглифическое письмо², разработана технология изготовления папируса и пергамента, сложилась практика изготовления манускриптов и т.д. Насколько известно это достаточно трудоемкие и дорогостоящие технологии. Например, изготовление одного пергаментного манускрипта требовало нескольких десятков голов скота и многих месяцев, а может быть и лет труда переписчиков. Очевидно, что столь высокие затраты могут быть рационально объяснены только укреплением и расширением органов власти. Горизонтальное «коммуникационное строительство» не могло позволить себе таких высоких затрат. Были созданы такие более дешевые и простые коммуникационные технологии, как фонетическое письмо, изобретены технология использования бумаги и книгопечатание, которые в последствии вытеснили более дорогие и трудоемкие процессы. Безусловно, это достаточно схематичное и условное видение эволюции коммуникационных

¹ Документная коммуникация, отмечает Н.С. Ларьков, представляет собой процесс движения в обществе документированной информации, т.е. информации, закрепленной на материальном носителе [13, с.338].

² Напомним, что буквально слово «иероглиф» обозначает священное письмо.

технологий, однако нам интересна сама тенденция, которая, по нашему мнению, просматривается.

Несмотря на свою значимость, коммуникационные технологии в контексте идей Лумана все же не более, чем средства преодоления физического пространства и времени. Непосредственное «коммуникационное строительство» состоит в образовании определенных «областей», позволяющих выборочно воспринимать и перерабатывать сложность (комплексность) мира, превышающих способности восприятия отдельного человека путем перераспределения знания и незнания. Что же это за области? Луман относит к ним науку, политику, право и т.д. и рассматривает их достаточно широко. Возможно это так, однако нам представляется, что помимо них существуют определенные промежуточные области, выполняющие функцию предварительного отбора сообщений, их аккумуляции с целью повышения качества «строительного материала». Это своего рода «обогачительные фабрики», которые не просто повышают выход металла из руды, но и в ряде случаев делают выплавку металла вообще возможной. При этом, чем сложнее конструкция «возводимой» социальной структуры, тем выше требования к отбираемым сообщениям, а соответственно, и к структурам, обеспечивающим эту процедуру. Каковы же требования к процедуре информационного «обогащения» сообщений? Во-первых, это касается информативности сообщений, что достигается их предварительной тематической селекцией и последующей тематической аккумуляцией в фондах библиотек, архивов или иных информационных учреждений. Традиционно эта процедура определяется как комплектование. Во-вторых, помещение сообщения в фонд гарантированно обеспечивает семантическую стабильность сообщения, что создает условия для его последующего комментирования. Это, как мы понимаем, также способствует повышению информативности. В итоге сообщение, как говорят культурологи, становится текстом. Именно комментарий, по мнению известного немецкого культуролога и египтолога Яна Ассмана, делает текст *текстом* [15, с.191]. В-третьих, это создание условий для последующего поиска и организации использования письменных сообщений (систематизация, каталогизация, библиографирование). И, наконец, в-четвертых, помещение сообщений в фонд обеспечивает необходимый уровень семантической симметрии в ходе их пространственно-временного движения. В совокупности все это обеспечивает необходимый уровень восприятия информации и ее последующее использование в конструировании социальных систем. И, как следствие, значимые управленческие, политические, экономические, религиозные и иные события не просто воссоздаются на бумаге, а в ряде случаев они становятся возможными, если только получают письменное оформление. Именно последнее стало причиной появления феномена подлога сообщения, когда содержание события, либо вообще факт его существования оказалось возможным создать искусственно, в своих интересах при помощи пера и бумаги. В силу этого оперирование «письменными» событиями оценивается как стратегический ресурс, обеспечивающий процессы управления в самом широком смысле.

Механизмом развития описанных нами выше промежуточных коммуникационных областей выступила их институционализация, в ходе которой они приняли форму библиотечно-библиографических, архивных, нотариально-делопроизводственных и иных учреждений. Каков статус этих учреждений? Как мы уже отмечали, Луман относит информацию к содержанию коммуникационного процесса. Нам представляется это справедливым лишь для части процессов, связанных с передачей сообщений. В том случае, когда имеет место информационное «обогащение» сообщений посредством их целенаправленного отбора, селекции, фондирования, комментирования, библиографического описания и ряда других, мы имеем дело с информационными процессами. Отсюда мы полагаем, что библиотечно-библиографические и архивные институциональные структуры носят комплексный коммуникативно-информационный характер. Сообщения, продуцируемые этими социальными институтами, определяются как документы, а действующие учреждения и системы – как документальные [16]. При этом ряд из них, в первую очередь нотариально-делопроизводственные, реализуют нормативно-методическую коммуникацию, целью которой выступает разрешение тех или иных действий, а если смотреть шире, то событий и порядка их реализации, либо запрещения. Когнитивная коммуникация, реализуемая библиотечно-библиографическими учреждениями, направлена на перераспределение знания и незнания. В архивах учреждения, регистратурах и ведомственных архивах доминирует коммуникация нормативно-методического типа, тогда как в исторических архивах – коммуникация когнитивного типа. Такой же комплексный характер присущ коммуникациям, реализуемым научно-информационными учреждениями (органами научно-технической информации). Что касается письменных сообщений, продуцируемых вне этих учреждений и систем, то они рассматриваются нами как продокументные [17].

В общей схеме эволюции социальных коммуникаций уточнялись роль и место документально-информационных институтов. Традиционно в этой схеме выделялось создание письменности и следом за ней – изобретение книгопечатания. Этой же схемы придерживается и Луман, который отмечает, что именно письменность и книгопечатание инициируют различия, конструирующие коммуникацию [2, с. 223]. Развивая эти идеи, А. Назарчук [18] отмечает, что письменность создала эффективный механизм исторической памяти и позволила расширить пространственные границы в повседневных коммуникациях, однако при этом, во-первых, сохранялась необходимость личного участия коммуницирующих людей, а во-вторых, – в намерение пишущего не входила публикация письма или книги, и, как следствие, количество публично доступных текстов было ничтожно мало. Кардинально ситуация, по мнению А. Назарчука, поменялась лишь с изобретением книгопечатания, приведшего к созданию «удаленной» коммуникации – коммуникации отсутствующих. В религиозной плоскости это привело к смене «мемориального типа чтения для воспоминания важных истин сакральных текстов экстенсивным типом, при котором в мате-

риале для чтения ищется новая информация» [18, с. 89]. Изобретение книгопечатания, по мнению Назарчука, привело к формированию новой коммуникативной среды, обеспечивающей накопление и трансляцию знаний, выработанных человечеством, открыло путь к формированию библиотек, а также дало возможность сортировать, сравнивать, классифицировать массивы знаний [18, с. 90].

Но так ли это, если включить в поле зрения документально-информационные учреждения и выполняемые ими коммуникативно-информационные функции? Действительно, как отмечал Назарчук, удешевление производства книжной продукции благодаря изобретению печатного станка, стимулировало коммуникацию и как следствие – социальное развитие. Это верно, но такое развитие было направлено, прежде всего, на горизонтальное конструирование социальных структур, что на практике проявилось в росте публичных библиотек. Однако сам феномен библиотеки как коммуникативно-информационного социального института возник гораздо раньше в системе вертикального социального конструирования и был нацелен на обеспечение развития вертикали власти. После изобретения книгопечатания библиотечно-библиографический институт оказался вовлечен в процесс горизонтального конструирования и таким образом претерпел лишь количественные, а не качественные изменения. Более того, наличие дворцовых, монастырских и университетских библиотек, располагающих фондами рукописных книг, обеспечило книгопечатанию источниковую базу и, как следствие, – ту высокую социальную значимость, которую оно получило впоследствии. Таким образом, тезис о существенном влиянии книгопечатания на библиотечное дело справедлив лишь отчасти.

ЗАКЛЮЧЕНИЕ

Подводя краткие итоги анализа теоретико-методологического потенциала концепции Никласа Лумана можно сказать, что его концепция направлена на раскрытие коммуникационных и информационных аспектов организации социума. По сложившейся в западной науке традиции информационные аспекты часто трактуются как составная часть коммуникационного процесса, что, наверное, оправдано теми задачами, которые стоят перед социологами. Однако в рамках информатики, нам представляется, что понятие информация и коммуникация, а также соответствующие им теоретические построения до сих пор сохраняют относительную научную автономность. Содержательно концепция Лумана носит семантико-прагматический характер, что открывает дорогу к более тесной интеграции теорий информации и коммуникации, существенно повышая их эвристический потенциал.

Что касается документально-информационных дисциплин, то, как нам представляется, наибольший интерес представляют идеи Лумана о роли и месте коммуникаций и соответствующих технологий в развитии социума. Рассматривая коммуникацию и ее эволюцию, Луман ограничивается анализом феноменов письменности и книгопечатания. Предметный анализ и развитие этих представлений позволяют рас-

ширить картину социальных коммуникативных структур, вводя в нее социальные документально-информационные институты, представленные библиотечно-библиографическими, архивными и нотариально-делопроизводственными учреждениями. До сих пор библиотечно-библиографические учреждения рассматриваются преимущественно с синтаксической точки зрения, т.е. как аккумулирующие письменные сообщения и предоставляющие их пользователям. В ряде случаев о них говорят как об институтах памяти. В контексте Лумана мы можем рассматривать деятельность библиотек в семантическом и прагматическом аспектах, что заключается в повышении информационной емкости сообщений, в стабилизации содержания сообщений и тем самым в повышении к ним доверия, в усилении вариативности и, как следствие, в повышении качества последующего процесса селекции сообщений и т.д. Таким образом, благодаря идеям Лумана библиотека вместе с другими социальными документально-информационными институтами предстает в новом свете, что позволяет нам расширить схему эволюции общества, включив в нее, помимо письменности, книгопечатания и других технологий, также и документально-информационные институты как структуры, обеспечивающие автопоэзис общества на качественно новом уровне.

Осмысление этого позволяет прийти к выводу, что концепция Лумана содержит в себе значимые методологические положения, которые обогащают наши представления о библиотечно-библиографических и иных документально-информационных институтах.

СПИСОК ЛИТЕРАТУРЫ

1. Луман Н. Что такое коммуникация? // Социологический журнал / пер. с нем. Д. В. Озирченко. – 1995. – № 3. – С. 114-125.
2. Луман Н. Социальные системы. Очерк общей теории / пер. с нем. И.Д. Газиева; под ред. Н. А. Головина. – СПб.: Наука, 2007. – 648 с.
3. Головин Н.А. Социология Н. Лумана – этап развития общей теории социальных систем // Социс. – 2014. – №2. – С. 86-94.
4. Назарчук А.В. Общество как коммуникация в трудах Никласа Лумана // Вопросы философии. – 2006. – №6. – С. 156-173
5. Назарчук А.В. Учение Никласа Лумана о коммуникации. – М.: Весь мир, 2012. – 246 с.
6. Антоновский А.Ю. Никлас Луман: эпистемологическое введение в теорию социальных систем. – М.: ИФРАН, 2007. – 134 с.
7. Ивахненко Е.Н. От рациональной коммуникации к коммуникативной рациональности // Вестник РГГУ. Серия: Философия. Социология. Искусствоведение. – 2014. – №10. – С. 97-105.
8. Посконина О.В. Никлас Луман о политической и юридической подсистемах общества. – Ижевск: Изд-во Удмуртского ун-та, 1997. – 122 с.
9. Плешкевич Е.А. Дискуссия о природе информации и путях построения ее философской концепции (обзор) // Научно-техническая информация. Сер.1. – 2015. – № 4. – С. 14-18;

- Pleshkevich E.A. A Discussion of the Nature of Information and Methods of Building Its Philosophical Concept (Review) // Scientific and Technical Information Processing. – 2015. – Vol. 42, № 2. – P. 53-58.
10. Соколов А.В. Введение в теорию социальной коммуникации: учеб. пособие. – СПб., 1996. – 319 с.
 11. Вохрышева М.Г. Теория библиографии : учеб. пособ. – Самара : Изд-во СГАКИ, 2004. – 368 с.
 12. Коршунов О.П. Библиографоведение : учебник. – М. : Книжная палата, 1990. – 232 с.
 13. Ларьков Н.С. Документоведение : учеб. – 3-е изд., перераб. и доп. – М. : Проспект, 2016. – 416 с.
 14. Фокеев В.А. Природа библиографического знания. – М., 1995. – 352 с.
 15. Ассман Я. Культурная память: Письмо, память о прошлом и политическая идентичность в высоких культурах древности / пер. с нем. М.М. Сокольский – М. : Языки славянской культуры, 2004. – 368 с.
 16. Плешкевич Е.А. Документальность как атрибутивное свойство документа // Научно-техническая информация. Сер.1. – 2013. – №9. – С.1-7; Pleshkevich E.A. Documentality as a Document's Attribute Feature // Scientific and Technical Information Processing. – 2013. – Vol. 40, № 3. – P. 146-151.
 17. Плешкевич Е.А. Протодокументные коммуникации: новая категория в коммуникационных технологиях // Научные и технические библиотеки. 2006.– №12. – С. 13-23.
 18. Назарчук А.В. Влияние книгопечатания на развитие протестантизма в Европе // Новая и новейшая история. – 2006. – №4. – С. 71-90.

Материал поступил в редакцию 14.04.16.

Сведения об авторе

ПЛЕШКЕВИЧ Евгений Александрович – доктор педагогических наук, Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук, главный научный сотрудник (Новосибирск)
e-mail: eap1966eap@mail.ru

П.А. Калачихин

Сервисно-ориентированная архитектура наукометрической информационной системы*

Рассматриваются вопросы проектирования государственной автоматизированной наукометрической информационной системы. После построения концептуальной модели на основании аналогий с другими программными средствами производится анализ опыта создания подобных систем в Российской Федерации и за рубежом. Предлагается оригинальный вариант реализации программной архитектуры, главной особенностью которого является возможность получения всех необходимых данных и предоставления имеющихся функций конечным пользователям через импорт и экспорт web-сервисов.

Ключевые слова: научные информационные системы, оценка наукометрических показателей, сервисно-ориентированная архитектура, управление наукой, web-сервисы

ВВЕДЕНИЕ

Рост количества научных публикаций в середине XX в. заставил обратить внимание на изменение характера научных исследований, в том числе на изменение значимости публикационной активности и информационной деятельности в работе каждого научного сотрудника. Именно в связи с этим внимание исследователей оказалось сосредоточено на тех показателях научной деятельности, которые могли быть оценены количественно и учтены автоматически в создаваемых банках данных – на количестве публикаций ученого и цитирований его работ, содержащихся в работах других ученых. Первый показатель довольно быстро обнаружил свою зависимость, в первую очередь, от внешних по отношению к исследовательскому процессу факторов – социальных и психологических, и в связи с этим на время перестал быть интересен для изучения процессов развития научного знания. Второй показатель оказался гораздо интереснее.

Перечень библиографических ссылок, указанных в публикации, создает своеобразный контекст работы, дает первое представление о проблемах, которые рассматриваются, является ключом к пониманию заложенных в ней идей. Исследователи науки исходят из предположения, что наиболее зримо процесс научной коммуникации представлен в журнальных публикациях. Выступления на конференциях и семинарах со временем выливаются в статьи, а монография

создается как итог работы, отраженной ранее с той или иной степенью полноты в статьях. С другой стороны, и для самого ученого статья – это итог изучения некоторого предмета, поскольку следующее его исследование, даже являющееся продолжением данного, потребует уже обновления методов и, соответственно, переопределения предмета. Таким образом, статья, рассматриваемая как знак в системе науки, представляет собой «единство общения и обобщения» и может считаться единицей анализа научной деятельности [1].

Цель нашего исследования – предложить вариант проекта Государственной автоматизированной наукометрической информационной системы (ГАНИС), предназначенной для оценки наукометрических показателей, характеризующих продуктивность работы ученых, и для оценки набора наукометрических показателей других вовлеченных в научную деятельность субъектов – научных организаций, фондов, изданий и пр.

Разработка наукометрических информационных систем – это сложная научная и инженерная задача, поэтому изложенные в настоящей статье положения ограничиваются лишь проектированием архитектуры государственной наукометрической системы с детализацией на уровне технического проекта. В связи с допущением многовариантных научно-методических и практических подходов к разработке и реализации проектов по созданию наукометрических информационных систем, необходимо сделать акцент на том, что предлагаемый в настоящей статье один из альтернативных вариантов архитектуры может и должен быть подвергнут обсуждению и критике, прежде чем

* Работа выполнена в рамках исследования по теме 0003-2015-0008 Госзадания ВИНТИ РАН

стать основой для реальных разработок в случае успешного прохождения стадии обсуждения и принятия заказчиками соответствующих решений по ее реализации.

Заказчиками создания Государственной автоматизированной наукометрической информационной системы выступают государственные ведомства – Министерство образования и науки и ФАНО (Федеральное агентство научных организаций). ГАНИС предназначается для ученых (аспиранты, докторанты, преподаватели, научные сотрудники), научных организаций (вузы, научно-исследовательские институты, фонды), чиновников в сфере науки и образования (потребители и производители сводной и отчетной информации), печатных изданий (журналы, издательства, и т.д.), а также для других субъектов, вовлеченных в научно-исследовательскую деятельность.

Проектирование архитектуры ГАНИС требует проведения предварительного анализа имеющегося отечественного и зарубежного опыта по созданию систем подобного класса. При этом необходимо иметь четкое объяснение, для чего и с какой целью будет создаваться система, в чем ее преимущества перед существующими аналогами. Проектирование архитектуры ГАНИС должно отвечать общим принципам ее построения. В нашей работе предлагается вариант архитектуры, основанный на использовании сервисов для извлечения исходных данных и для подключения к ГАНИС с целью ее использования. Проектные решения по созданию ГАНИС должны отвечать экономическим интересам государства, в частности целям инновационного развития экономики, сильно зависящей от эффективности научной деятельности. В таком ключе главной задачей создания и внедрения ГАНИС является улучшение качества обслуживания научно-исследовательских, а вместе с ними и научно-инновационных процессов в различных сферах экономики Российской Федерации, нуждающихся в инновациях. В свете изложенного ГАНИС берет на себя роль своеобразной платформы, не участвующей в исследованиях на переднем крае науки, но являющейся надежной точкой опоры для отечественных исследователей, организаций и других субъектов, заинтересованных в научно-инновационной деятельности.

КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ ГАНИС

Построение и функциональные возможности базы данных цитирований публикаций для японских академических журналов и их импакт-факторов описаны в [2].

База данных цитирований японских публикаций обслуживается с помощью интерфейса доступа, предоставляющего пользователям возможности осуществлять объединенные поиски по всем базам данных. База данных цитирований японских публикаций уникальна по сравнению с другими библиографическими базами данных, поскольку она держит данные о цитатах. Поиск дает список соответствующих публикаций, и пользователи могут получить библиографические данные и информацию о цитатах для каждой публикации. В ходе развития этой автоматизированной системы были вычислены импакт-факторы для

японских журналов, включенных в базу данных цитирований японских публикаций. База данных цитирований японских публикаций используется как основная база данных для того, чтобы координировать научные публикации, авторов, исследователей, проекты и научные организации через цитаты. В установлении таких связей объединение библиографических данных и отчетов по цитатам требует нечеткого соответствия для того, чтобы принять изменения в описаниях и объединить их. Из-за длительных экономических трудностей японские предприниматели очень заинтересованы в оценке исследований для эффективных инвестиций в научно-исследовательские работы, которые принесли бы определенное технологическое развитие и усилили бы национальную конкурентоспособность на международном рынке. Наукометрическая система, как ожидается, будет работать на эту цель двумя различными способами следующим образом. Во-первых, она ускорит процесс исследования, эффективно предоставляя информацию разработчикам. Во-вторых, она обеспечит некоторые меры по оценке, которые принудят руководителей исследований в национальных правительствах и местных органах власти, университетах, научно-исследовательских институтах и промышленности эффективнее выстраивать приоритеты для предполагаемых областей экономики.

Созданная в Российской Федерации система автоматизации библиотек (САБ) «Ирбис 64» обладает возможностью загрузки в базу данных готовых записей из индексов цитирования. Возможность выгрузки записей есть в нескольких индексах цитирования, в случае Web of Science и Scopus требуется наличие подписки на эти ресурсы, а в случае РИНЦ (Российский индекс научного цитирования) необходима подписка на информационно-аналитическую систему Science Index. Из Web of Science и Scopus можно выгружать любые записи, ограничение существует только на разовую порцию выгрузки. Science Index допускает скачивание только тех записей, которые приписаны к организации, имеющей подписку. Из Web of Science и Scopus возможен экспорт записей в текстовом формате, а также определение набора элементов выгружаемых записей. В Science Index вероятно выгрузка в формате XML полного набора всех элементов записи. Web of Science и Science Index предоставляют при выгрузке идентификаторы своих записей. В записях, экспортированных из Scopus, идентификаторов нет. Для загрузки в БД САБ «Ирбис 64» данных, выгруженных в текстовом формате, используется специальный конвертор.

После первичного отбора и загрузки записей за весь доступный период для получения информации о появлении новых публикаций в Web of Science и Scopus выполняются постоянно действующие запросы, информирование о новых поступлениях осуществляется еженедельно, с этой же периодичностью происходит пополнение базы данных публикаций САБ «Ирбис 64». Экспорт информации из индексов цитирования, конвертирование и загрузка в базу данных выполняются библиотекарем.

Получение актуальных данных о количестве цитирований публикаций в Web of Science и Scopus

полностью автоматизировано. Для поддержания информации о цитировании в актуальном состоянии используются пакетные задания. Каждое пакетное задание последовательно выполняет три задачи:

- выгрузка идентификаторов из базы «Ирбис 64»;
- получение данных о количестве цитирований из базы Scopus или Web of Science;
- обновление данных в базе «Ирбис 64».

Обмен данными между задачами осуществляется посредством текстовых файлов специального формата. Далее для получения данных о количестве цитирований по базам Scopus и Web of Science используется специально разработанное отдельное консольное приложение. Основными задачами приложения являются чтение входного файла с идентификаторами, выполнение запросов к API (Application Programming Interface), формирование выходного файла, дополненное информацией о количестве цитирований данной статьи. Затем средствами импорта системы «Ирбис 64» эта информация загружается в базу данных публикаций сотрудников [3].

Система «Соционет» использует принципы и стандарты технологии OAI (Open Archives Initiative – инициатива открытых архивов) и расширяет традиционные функциональные возможности базирующихся на них открытых архивов за счет развитых средств управления содержанием архива, а также средств создания связей между информационными объектами как внутри архива, так и с объектами из других архивов. Инициатива OAI – это подход к созданию распределенных электронных репозиториях информационных ресурсов. Ключевой для этого подхода является концепция открытого архива. Открытый архив – это распределенная совокупность доступных в среде Интернет уникальным образом идентифицируемых информационных объектов, для которой поддерживается репозиторий описывающих их стандартизованных метаданных, представляемых средствами языка разметки XML. Репозитории метаданных открытых архивов должны быть доступны для авторизованных сборщиков метаданных, функционирующих в соответствии с разработанным протоколом. Для этого каждый репозиторий располагает механизмом доступа, способным выполнять по запросам сборщиков метаданных набор операций, определенных протоколом OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting).

Благодаря стандартизации представления метаданных в репозиториях и протоколе OAI-PMH, определяющему взаимодействие открытых архивов, обеспечивается интероперабельность открытых архивов. Согласно протоколу OAI-PMH, для заданного открытого архива может осуществляться сбор метаданных из репозитория метаданных зарегистрированных в нем других открытых архивов в соответствии с установленным регламентом, а также синхронизация состояния репозитория метаданных рассматриваемого открытого архива с изменениями состояний репозитория зарегистрированных в нем архивов – источников метаданных.

В результате такой открытый архив выступает в качестве своеобразного информационного «хаба» – системы интеграции данных, основанной на мате-

риализованной интеграции метаданных из зарегистрированных в нем источников – открытых архивов. Описываемые метаданными информационные объекты-публикации при этом никак не затрагиваются. Они подвергаются виртуальной интеграции. В интегрированном репозитории метаданных обеспечивается их уникальная идентификация. Полные тексты этих публикаций сохраняются в прежних местах хранения и становятся доступными с помощью пользовательских сервисов рассматриваемого интегрированного открытого архива по ссылке на него в описывающей этот объект карточке метаданных.

Организация «Соционет» следует описанным принципам. Интегрируемые открытые архивы – источники ее информационных ресурсов – распределены в сетевой среде и доступны в этой среде. Некоторые локальные репозитории метаданных могут размещаться на серверах самой системы в так называемых «личных зонах». После создания коллекции информационных объектов в локальном репозитории администратор этой коллекции включает так называемый режим трансляции, после чего эта коллекция становится доступной для сборщика метаданных и включается в процедуру сбора метаданных, осуществляемую сервисами системы «Соционет». В соответствии с установленным регламентом, системные сервисы выполняют синхронизацию состояния метаданных в централизованном глобальном системном репозитории с состоянием их во всех транслируемых коллекциях локальных репозиториях метаданных интегрируемых источников. Эта процедура реализуется, как уже отмечалось, в соответствии с протоколом OAI-PMH.

Система «Соционет» с ее централизованным интегрированным репозиторием метаданных, в свою очередь, может служить источником информационных ресурсов для других систем, представляющих собой открытые архивы OAI.

В качестве системы интеграции информационных ресурсов и открытого архива OAI система «Соционет» выполняет следующие функции:

- регистрация открытых архивов – источников информационных ресурсов для системы «Соционет», которыми могут выступать, в частности, и локальные репозитории метаданных, базирующиеся на серверах системы «Соционет» (в «личных зонах»);
- поддержка создания коллекций информационных объектов и описывающих их метаданных в «личной зоне» администратора коллекции;
- обеспечение уникальной идентификации информационных объектов в интегрированном репозитории метаданных системы;
- сбор метаданных из зарегистрированных открытых архивов и синхронизация состояний репозитория метаданных системы и репозитория метаданных открытых архивов – источников информационных ресурсов;
- обеспечение для пользователей «Соционет» навигации по информационному пространству системы, просмотра метаданных информационных объектов и доступа к их полному контенту, если он доступен;
- поиск информационных объектов по критериям, заданным поисковыми сервисами «Соционет».

Оценивая «Соционет» как систему интеграции информационных ресурсов множества открытых архивов, следует отметить, что она обладает дополнительными возможностями интеграции информационных ресурсов по сравнению с другими системами, основанными на технологиях ОАИ. Эти возможности обеспечиваются естественным образом как побочный эффект существования «Соционет» в сетевой среде. Способ интеграции информационных ресурсов, который имеется в виду, можно назвать навигационной интеграцией. Информационные объекты «Соционет» могут быть связаны с информационными объектами, доступными в сетевой среде, но не описанными в «Соционет» и тем самым «не принадлежащими» ресурсам системы. Например, в карточке метаданных публикации может быть гиперссылка на интернет-страницу, содержащую рецензию на эту публикацию. В свою очередь, на этой странице могут иметься ссылки на другие страницы. Такие внешние (по отношению к ресурсам «Соционет») связи позволяют пользователю осуществлять навигацию не только по структуре информационного пространства «Соционет», но и «бесшовно» выходить за его пределы в сетевую среду и возвращаться обратно [4].

РИНЦ и база данных системы eLibrary различаются, поскольку последняя, помимо журналов, входящих в РИНЦ, содержит: иные русскоязычные журналы (включая научно-популярные, рекламно-информационные); неперiodические издания на различных языках (монографии, учебники, пособия и т.п.); зарубежные журналы на иностранных языках, издатели которых в обозримом будущем не планируют процедуру вхождения в РИНЦ. При этом eLibrary учитывает и ненаучную литературу. Хотя ее относительная доля может быть невелика, но известны примеры изменения индекса цитируемости в десятки и даже сотни раз.

Системы Web of Science и Scopus построены на данных жестко ограниченного списка изданий. Если ученый имеет некоторый рейтинг, то мы знаем, сколько он имеет публикаций в журналах из строго определенного перечня. В случае eLibrary на главной странице поиска «Авторский указатель» выдается общее число проиндексированных документов и ссылок на них из любых источников. В результате дополнительных запросов можно получить данные по РИНЦ и списку ВАК, но в этом случае будут потеряны данные зарубежных журналов.

Большой ошибкой является включение монографий, пособий, разовых сборников (фактически, любых текстов) в базу для расчета количества публикаций, цитирования, индекса Хирша. Такой подход противоречит международной практике.

Качество результата запроса в eLibrary, в первую очередь, определяется наличием регистрации ученого в этой системе и тем, насколько оперативно и добросовестно отнесся он к процедуре выбора своих работ и ссылок на них. Появление eLibrary имеет объективные основания, связанные с традицией писать научные тексты на родном языке. Аналогичные системы развивают в ряде зарубежных стран.

Система eLibrary прошла стадию первичного накопления данных и можно начинать практику ее

применения для получения наукометрических характеристик как ученых, так и отечественных журналов. Использование международных баз данных для определения наукометрических показателей в наибольшей степени способствует сохранению потенциала отечественной науки. Однако, поскольку Web of Science и Scopus охватывают публикации лишь 2–10 % научных работников, то создание национальной базы данных РИНЦ на основе eLibrary, несмотря на все проблемы, связанные с неполнотой и способом формирования данных (индексирование из любых источников), является важным положительным шагом в сохранении российской науки [5].

Наиболее значимыми считаются статьи, опубликованные в журналах с более высоким импакт-фактором, а также индексируемые признанными международными системами – Scopus, Web of Science, Web of Knowledge, Springer и др. По состоянию на текущий день активно ведется работа по формированию отечественной системы РИНЦ, которая пока, к сожалению, не располагает достаточно представительным корпусом публикаций для наукометрических измерений, но охватываемое ею научное публикационное пространство интенсивно развивается.

Представление в научной электронной библиотеке семантических связей как самостоятельных информационных объектов обеспечивает ряд преимуществ по сравнению со «встроенным» способом их представления. Действительно, на основе контента электронной библиотеки может быть создан автономный по отношению к ее контенту репозиторий семантических связей, который может служить новым источником данных для наукометрии. Такие репозитории, созданные в различных библиотеках, могут интегрироваться и благодаря этому формировать глобальные репозитории семантических связей для той или иной области знания. На этом пути возможно создание достаточно представительных источников данных для наукометрии в различных областях науки. Если электронные библиотеки построены на основе технологии открытых архивов ОАИ, то легко обеспечивается интеграция созданных в их среде репозитория семантических связей контента самих электронных библиотек.

Эта деятельность дополняет традиционную практику анонимного рецензирования печатных изданий, а открытость для научного сообщества высказанных оценок позволяет ответным образом реагировать на них, способствует более высокой ответственности и объективности их авторов.

Одним из главных результатов проекта Common European Research Information Standart (CERIF) является создание унифицированной концептуальной схемы, называемой «полной моделью данных». Эта модель рассматривается как единая основа создания научных информационных систем (НИС) для поддержки научно-организационной деятельности в разных странах и различных научных организациях. Благодаря стандартизации концептуальной схемы обеспечивается интероперабельность таких систем. В последнее время в рамках проекта CERIF была предложена спецификация стандартизованной семантики полной модели данных, а также онтология, опреде-

ляющая систему терминов для обозначения сущностей этой модели и отношений между ними.

В электронной библиотеке можно поддерживать несколько таксономий связей, основанных на разных онтологиях, точно так же, как для рубрикации информационных объектов можно использовать несколько рубрикаторов научно-технической информации.

Интеграция ресурсов семантических связей различных научных электронных библиотек может быть осуществлена известными методами виртуальной или материализованной интеграции данных из множества информационных источников. Интеграция относительно легко реализуется при условии базирования библиотек-источников на технологии ОАИ и протоколе ОАИ-РМН [6].

Используя компьютерную терминологию, будет правильно утверждать, что разработка ГАНИС является IT-проектом по созданию наукометрической информационной системы, в начале которого необходимо составить концепцию и написать техническое задание, опираясь при этом на имеющиеся разработки в виде прототипа грантовой наукометрической системы. При разработке ГАНИС, целесообразно воспользоваться зарубежным опытом создания государственных наукометрических систем. Однако заранее следует указать на то, что ГАНИС должна строиться как открытая система на основании профильных стандартов OSI RM и OSE RM [7, 8].

ГАНИС следует рассматривать как подвид наукометрических информационных систем. В свою очередь, наукометрические информационные системы являются одним из подклассов научных информационных систем (НИС) или разновидностью библиотечных информационных систем (БИС). НИС – это система, автоматизирующая процессы научного информационного обмена [9].

ПРЕДПРОЕКТНОЕ ОБСЛЕДОВАНИЕ

До недавнего времени в российской научно-исследовательской деятельности преобладала территориально-распределенная межведомственная информационная система с «медленными» сервисами. «Медленными» значит частично автоматизированными, т. е. научные организации были вынуждены обращаться в эту систему с запросами для предоставления им наукометрической информации, которая в достаточно серьезном объеме обрабатывалась ручным способом без полноценных инструментов автоматизации данного процесса. Таким образом, в качестве модели «AS-IS» имеется информационная система, обрабатывающая документопоток далеко не самым эффективным способом, а в качестве модели «TO-BE» должна быть легковесная и доступная программа или платформа, выполняющая все необходимые вычисления и манипуляции с данными прямо «на лету».

В качестве реального примера следует привести отечественную Автоматизированную информационно-библиотечную систему по общественным наукам (АБИС) Института научной информации по общественным наукам (ИНИОН) РАН. В ИНИОН была предложена система научной информации в виде трех ступеней: библиографической (система БУ), ре-

феративной (система РЖ) и аналитической информации (научно-аналитические обзоры литературы по основным проблемам и отраслям общественных наук, по отдельным странам и регионам мира, имеющие прогностическую функцию и функцию моделирования социального развития). Сегодня в составе ИНИОН действует ряд специализированных центров научно-информационных исследований, фундаментальная библиотека, центр информатизации, издательский центр [10].

Другим примером, в который заложен принципиально иной подход, является информационно-поисковая система (ИПС) «Мимоза», которая поставляется Роспатентом вместе с библиографическими данными и рефератами патентов на DVD-диске. Для обеспечения дополнительного структурирования и обработки электронных форм описаний изобретений был разработан комплекс программ в Институте проблем управления (ИПУ) РАН.

DVD-диск содержит ссылки на сайт Роспатента в сети Интернет, обеспечивающие бесплатный доступ к полным описаниям изобретений. Таким образом, пользователь ИПС «Мимоза» после выполнения поиска на DVD-диске библиографических данных и рефератов нужных ему изобретений с помощью компьютера, подключенного к сети Интернет, сразу же может получить доступ к электронным формам полнотекстовых описаний изобретений по ссылкам на соответствующие страницы сайта Роспатента.

Используя DVD-диск, с помощью ИПС «Мимоза» пользователь по заданному запросу выбирает библиографические данные и рефераты описаний изобретений (например, по автору изобретения, дате публикации или по коду Международной патентной классификации – МПК). По ссылкам на сайт Роспатента выбираются электронные формы полнотекстовых описаний изобретений, соответствующие найденным библиографическим данным и рефератам описаний изобретений. Из каждого описания изобретения автоматически копируется часть библиографических данных (код МПК, название, автор изобретения и другие поля, необходимые для вычисления индикаторов). С помощью разработанного в ИПУ РАН комплекса программ в каждом найденном описании изобретения выделяются, копируются и сохраняются ссылки на цитируемые документы. При этом для каждой ссылки указываются вид публикации и фрагмент в документе, где встретились ссылки (поле библиографических данных, абзацы текста описания или список использованной литературы в конце текста), а для ссылок из текста – адрес в тексте описания. Текст каждой ссылки на цитируемый документ структурируется. Задавая тематические запросы, можно формировать массивы структурированных ссылок на цитируемые документы, соответствующие заданному запросу [11].

ПРОГРАММНАЯ АРХИТЕКТУРА ГАНИС

Все электронные публикации – это некоторые файлы, представленные в необходимом формате. С точки зрения использования системы, форматы подразделяются на входные (авторские) и выходные (читательские). Входные форматы являются либо

общеиспользуемыми (Word, TeX, HTML), либо открытыми (DocBook, TEI), основанными на XML. Читатель будет иметь возможность получить электронную публикацию в зависимости от его требований: HTML, DjVu, LuraDocument – для ознакомления; PDF, PS – для печати; WML – для беспроводных устройств; eBook (OpenBook) – для чтения в электронных книгах; подмножество XML – для обмена; исходные Word, TeX – для дальнейшей работы. При этом получаемый документ может быть адаптирован под физические данные читателя или программные и аппаратные возможности его системы [9].

Для представления данных о научных исследованиях в виде метаданных необходимы: схема метаданных – формальное определение структурного представления информационных ресурсов и связей между ними; словари – списки названий, организованные иерархически либо в виде простого перечисления для придания точного смысла (семантики) названиям информационных ресурсов. Например, CERIF описывает основные типы научных данных. В настоящее время разработан ряд БД на базе этого стандарта, а также разработана схема для кодирования данных CERIF в RDF и технологии для экспорта и импорта RDF из базы данных CERIF [12].

Механизм подписки на агрегированную информацию из наукометрических баз данных на основании web-сервисов является основным компонентом ГАНИС. Для этой цели должен быть реализован специальный интерфейс доступа к системе. В данном случае уместна аналогия с новостными агрегаторами, которой следует воспользоваться, чтобы описать, каким образом происходит сбор данных. Подключение к наукометрическим базам данных должно происходить через web-сервисный «пул», а выдача выходных результатов также должна осуществляться через набор web-сервисов.

Если бы пришлось отказаться от импорта данных через web-сервисы, то прямой ввод данных мог бы осуществляться через чтение файлов с публикациями

по URL, файловый загрузчик или с помощью интернет-поисковика. При этом было бы необходимо загружать не только научные публикации, но и библиографические данные. В любом случае, описание публикаций требует настройки библиографических форматов, поэтому вся библиографическая информация должна заноситься в специальные регистры. Кроме того, необходимо реализовать загрузку справочного файла со списком импакт-факторов журналов. Отсюда вытекает, что будет нужно поддерживать разные книжные форматы хранения данных.

Но, благодаря сервисно-ориентированной архитектуре, сканирование, распознавание и синтаксический разбор публикаций осуществлять не потребуется, в этом заключается главное преимущество предлагаемого решения. Поэтому нет необходимости оставлять полные тексты публикаций в хранилище данных. Интеграция с поисковыми системами может осуществляться через «шлюз», поддерживающий разные сетевые протоколы, поэтому очень важно будет реализовать возможность формирования отчетов по результатам поисковых запросов к наукометрической информации.

Схема потоков данных Государственной автоматизированной наукометрической информационной системы представлена на рис. 1.

С целью экономии ресурсных затрат при создании ГАНИС за основу целесообразно взять «Указатель РФФИ», созданный в 2002-2003 гг. и реализованный в сетевом варианте, поэтому доступ к нему возможен как через Интернет, так и через Интранет. «Указатель РФФИ» основан на идеях и технологии «Системы по использованию и анализу научной информации» 3.4 (СИАНИ), зарегистрированной в Роспатенте в 2002 г. под номером № 2002611423. Источником данных для «Указателя РФФИ» является БД РФФИ, в которой содержится информация о поданных на конкурсы РФФИ заявках на исследовательские проекты и результатах их выполнения.

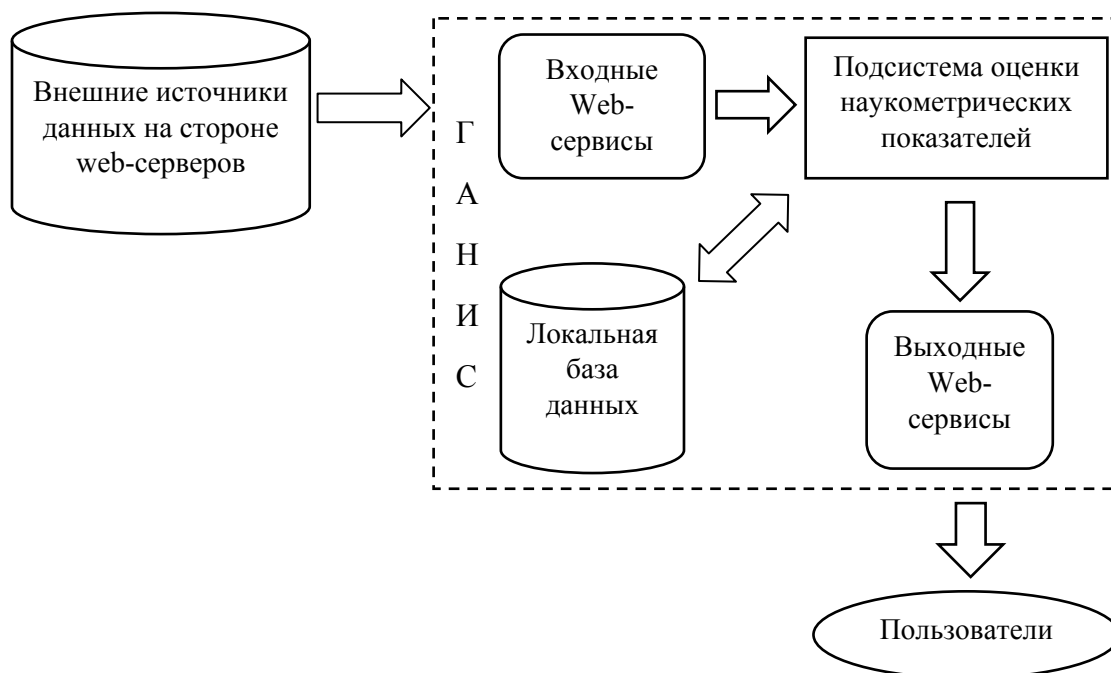


Рис. 1. Схема поступления данных в Государственную автоматизированную наукометрическую информационную систему

Функционально «Указатель РФФИ» включает:

- подсистему поиска – позволяет осуществлять поиск по каждому из типов информационных объектов и в каждом из нескольких десятков полей, ранжировать результаты поиска, вести протокол поиска. Объектами могут быть проект, отчет, организация, ведомство, индивидуальный исследователь, публикация, научный журнал;
- подсистему навигации – дает возможность на основании полученных на этапе поиска данных переходить к другим данным, связанным с исходными данными;
- указатель документального цитирования – функционально представляет собой аналог Science Citation Index;
- указатель объектного цитирования – регистрирует все обращения, содержащие формальные признаки запроса, имя пользователя, время обращения к информационным объектам – публикациям, проектам, отчетам, организациям и т.д.;
- подсистему связи – инструмент, позволяющий зарегистрированному пользователю «Указателя РФФИ» практически мгновенно связаться с любым другим зарегистрированным пользователем этого ресурса [13].

Для того чтобы четко представлять взаимосвязь объектов и, следовательно, в полной мере иметь возможность оперировать информацией, распишем эти взаимосвязи, исходя из парного принципа (взаимосвязь близлежащей пары информационных объектов):

♦ ведомство / организация – это организация, которая находится в ведении ведомства (или ведомство, которое включает организацию);

♦ организация / проект – это организация, которая является основным участником – исполнителем проекта (далее организацией финансирования);

♦ персона / организация – это персона (заявитель или фандрайзер), которая является работником данной организации;

♦ персона / проект – это персона (заявитель или фандрайзер), которая является участником проекта (заявки или гранта);

♦ проект / отчет – это отчет о научно-исследовательской работе по результатам выполненного проекта;

♦ отчет / публикация – это публикация, указанная в научном отчете;

♦ публикация / журнал – это журнал, в котором представлена публикация (только для журнальных статей).

Причина возникновения такого количества информационных объектов заключается в том, что информация, указанная в заявке и, тем более, научном отчете, более чем обширна. Сложность при работе с данными, представленными в виде объектов, возможно, поначалу будет затруднять работу с системой. Но, по мере освоения системы и приобретения необходимых навыков – в первую очередь, именно в освоении нового восприятия, пользователь почувствует себя более уверенно. Кроме того, можно будет на практике убедиться в том, что именно такое объектное представление является единственной возможностью для гибкого, всестороннего оперирования и изучения информации. Двусторонние связи между объектами информационной базы отображены на рис. 2.

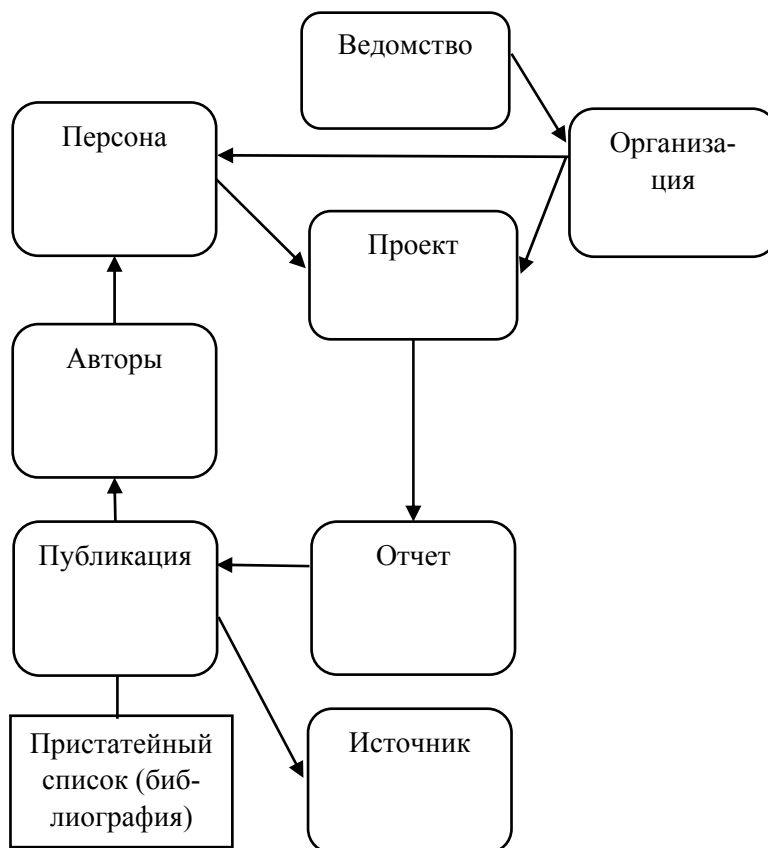


Рис. 2. Связи между объектами информационной базы Государственной автоматизированной наукометрической информационной системы

Среда Microsoft Access располагает мощным и гибким языком программирования и при этом не требует больших вычислительных ресурсов. Именно в этой среде реализован интерфейс базы данных «Указателя РФФИ», а сами данные хранятся в Microsoft SQL Server. Располагая web-интерфейсом для доступа к информационной системе, следовало бы учитывать все обращения пользователей и использовать данные, описывающие эти обращения, в качестве аналога литературных ссылок. При этом база данных должна быть в состоянии не только осуществлять авторизацию и идентификацию каждого обращающегося к системе пользователя, но и сохранять информацию о каждом его запросе, фиксировать, хранить и обрабатывать эти сведения. Особое значение приобретают решение проблемы несанкционированного доступа и реализация надежной политики информационной безопасности [14].

ГАНИС проектируется как «научомерический сервер», сервисы которого возможно использовать, подключившись к нему программно. Предлагаемая архитектура ГАНИС не является единственно возможной, может иметь место и совсем другая архитектура. Преимущество предложенной архитектуры в том, что она не требует распознавания оригинальных текстов статей. Недостаток заключается в зависимости такой ГАНИС от поставщиков информации, т. е. расположенных в сети хабов с наукомерическими данными.

ЗАКЛЮЧЕНИЕ

Техническая реализация проекта по созданию Государственной автоматизированной наукомерической информационной системы должна выполняться в рамках научно-исследовательских работ прикладного характера, в то время как тема нашего исследования носит скорее теоретический и методологический характер. Прежде чем приступить к разработке ГАНИС, следует сформировать четко обоснованное понимание того, зачем нужно тратить ценные ресурсы и время на создание новой информационной системы, когда на рынке уже существует достаточное количество подобных программ, платформ и баз данных (таких, как РИНЦ и Science Index), известных широкому кругу специалистов и при этом вполне удовлетворяющих имеющимся запросам. Вероятно, здесь имеет смысл не заниматься разработкой еще одного нового программного продукта, а осуществить интеграцию имеющихся решений в единый комплекс, попутно выполняя те или иные модификации. В таком случае, ключевым инструментом системной интеграции будет выступать механизм web-сервисов, с помощью которых будет возможно подключаться к внешним источникам данных в гетерогенной информационной среде и получать оттуда всю необходимую информацию, а затем передавать переработанные данные, предоставляя их другими web-сервисами через интерфейсы удаленного доступа.

С практической точки зрения намного удобнее будет использование прототипов уже имеющихся программных средств, в которых в достаточной мере реализованы функции оценки наукомерических показателей, что предпочтительнее, чем создание новых программных средств совершенно «с нуля». Исходя из

этих позиций, полученная в нашем исследовании модель сервисно-ориентированной архитектуры имеет ценность для дальнейших исследований в области проектирования и разработки для отечественной науки наукомерических информационных систем, действующих на государственном уровне.

СПИСОК ЛИТЕРАТУРЫ

1. Кузнецова Ю.М., Осипов Г.С., Чудова Н.В. Изучение положения дел в науке с помощью методов интеллектуального анализа текстов // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 106-138.
2. Negishi M., Sun Y., Shigi K. Citation Database for Japanese Papers: A new bibliometric tool for Japanese academic Society // Scientometrics. – 2004. – Vol. 60, Iss. 3. – P. 333–351.
3. Баженов С.Р., Данилин М.В., Рогозникова О.А. Интеграция базы данных публикаций организации с индексами научного цитирования: реализация средствами САБ ИРБИС64 // В сборнике: Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса Международный профессиональный форум: Книга. Культура. Образование. Инновации. – URL: <http://gpntb.ru/win/inter-events/crimea2015/disk/042.pdf>. «Крым-2015». 2015.
4. Коголовский М.Р., Паринов С.И. Информационные ресурсы, наукомерические показатели и показатели качества метаданных системы Соционет // 9-я Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Переславль-Залесский, 2007. – URL: <http://www.ipr-ras.ru/articles/koga-pari07-1.pdf>. – RCDL'2007.
5. Савельева Ю.В., Хоперсков А.В. Научные журналы и эффективность научной работы: поисковые системы и базы данных // Управление большими системами: сборник трудов. – 2013. – № 44. – С. 381-407.
6. Коголовский М.Р., Паринов С.И. Новый источник данных для наукомерических исследований // В сборнике: Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды XV Всероссийской научной конференции RCDL'2013. – М., 2013. – С. 107-117. – URL: http://rcdl.ru/doc/2013/paper/s9_1.pdf.
7. ISO 7498:1996, Information processing systems ISO 7498:1996, Information processing systems – Open Systems Interconnection – Basic Reference Model. – URL: [http://standards.iso.org/ittf/PubliclyAvailableStandards/s020269_ISO_IEC_7498-1_1994\(E\).zip](http://standards.iso.org/ittf/PubliclyAvailableStandards/s020269_ISO_IEC_7498-1_1994(E).zip).
8. ISO/IEC DTR 14252, Portable Operating System Interface for Computer Environments System Interface for Computer Environments – POSIX. – URL: http://www.iso.org/iso/catalogue_detail.htm?csnumber=23985.
9. Гребнев А.Н. Научные информационные системы // Вестник УдГУ: Математика. – Ижевск: Изд-во УдГУ, 2003. – С. 99–106.

10. Черный Ю.Ю., Соколова Н.Ю., Юрченкова Л.В. Отечественная информационно-библиотечная система по общественным наукам (1918-2009) // Теория и практика общественно-научной информации. – 2011. – № 20. – С. 121-139.
11. Зацман И.М., Шубников С.К. Принципы обработки информационных ресурсов для оценки инновационного потенциала направлений научных исследований // В сборнике: Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Девятой Всероссийской научной конференции RCDL'2007. – М., 2007. – С. 35-44. – URL: http://elibrary.ru/html_article.asp?id=22522202.
12. Лопатенко А.С. Современные научные информационные системы. Перспективы использования. – URL: http://e-science.narod.ru/CRIS_DOC.htm.
13. Либкинд А.Н., Минин В.А., Либкинд И.А., Маркусова В.А., Янсц М. Указатель РФФИ как инструмент информационного поиска и анализа тенденций в фундаментальной науке России // 7-я международная конференция «Информационное общество, интеллектуальная обработка информации, информационные технологии» – НТИ-2007, Москва, 24-26 октября 2007. – М.: ВИНТИ, 2007. – С. 180-183.
14. Алфимов М.В., Либкинд А.Н., Либкинд И.А., Минин В.А. Информационные потоки в РФФИ: новый подход к цитированию // Вестник РФФИ. – 2001. – № 4 (26). – С. 5-23.

Материал поступил в редакцию 01.07.16.

Сведения об авторе

КАЛАЧИХИН Павел Андреевич – кандидат экономических наук, старший научный сотрудник Отделения научных исследований по проблемам информатики ВИНТИ РАН, Москва
e-mail: pakalachikhin@viniti.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81' 32' 33

Н.В. Максимов, О.Л. Голицына, Е.И. Яковенко

Об одном подходе к оценке сложности словосочетаний и предложений на основе фрактальной аналогии

Предложена основанная на фрактальном подходе модель, позволяющая количественно оценить сложность текстовых конструкций с точки зрения фрагментарности представления смысла. По аналогии с фрактальной размерностью на основе системно-функционального подхода введены показатели размерности словосочетаний, обозначающих понятия, и предложений естественного языка. Представлена геометрическая иллюстрация векторной модели формирования смысла, выражаемого такими конструкциями, как словосочетания и предложения.

Ключевые слова: фрактальная размерность, смысловое значение, конструкция смысловыражающей единицы, размерность словосочетания, размерность предложения

ВВЕДЕНИЕ

Попытки выявления закономерностей в естественных языках (ЕЯ) – «внести регулярность в нерегулярное употребление языка» [1], в лингвистике представлены достаточно широко. Одно из таких направлений – разработка фрактальной модели языка. Систему единиц языка и речи можно представить в виде иерархической структуры: элементы одного уровня, объединяясь, становятся «кирпичиками» для построения более сложных образований, которые, в свою очередь, служат исходным материалом для конфигураций еще более сложного уровня и т.п. Слова, как минимальные единицы смысла и фрагменты текста, наделяют описываемый объект свойствами, комбинируются в словосочетания и далее в предложения, из которых формируется документ.

«Текст и его смысловое значение непрерывно связаны: текст состоит из фрагментированных единиц, которые при сочленении друг с другом образуют целостную, завершённую с точки зрения смысла, конструкцию. При этом в человеческой речи всегда одновременно присутствуют два компонента – видимый (или слышимый) текст и невидимый (и, соответственно, неслышимый) подтекст, который пребывает в душе человека. Порождение составных речевых единиц, начиная со сложных словосочетаний, предложений и заканчивая текстом, происходит путём линейного наращивания синтагм – конструкций, устойчивых с точки зрения смысла. Таким образом, конструкция

предложения в языке оказывается сложной и самоорганизующейся и не всегда подчиняется формальным правилам» [2].

«Языковые знаки разного уровня сложности входят один в другой и функционируют по принципам взаимовключения, самоидентичности в возрастающих масштабах. Это подтверждает фрактальное строение функционирующей языковой материи. Если согласиться с фрактальным строением языковой материи, то она состоит из взаимно идентичных элементов, наполняющих собой бесконечно расширяющийся дискурс. Если это так, то в языковой материи можно обнаружить фракталы концептуально-смыслового типа. ... Концептуально-смысловые фракталы образуют содержательную структуру текста, в котором элементы смысла объединяются в единое целое и обнаруживаются как на уровне слов в виде лексической связности текста целостного произведения, так и на уровне предложений, абзацев и пр. Данные фракталы обеспечивают содержательную целостность текста в целом, а также представляют собой референтную (служашую для связи) систему, соотносящую текст и его смысловое значение» [1].

Проявлением фрактальности считается и «размывание смысловых границ» между словом, словосочетанием и предложением [3].

Однако, как показывают и упомянутые работы, в лингвистике в основном рассматриваются качественные свойства, не позволяющие ввести количествен-

ную меру. В настоящей работе предложен подход, учитывающий иерархию конструкций и влияние частей речи на сложность текста. Фрактальность текстовых конструкций рассматривается с точки зрения точности описания предмета внимания (объекта, процесса, свойств) в контексте задачи его воспроизводства по этому описанию. При этом чем детальнее текст (и фрагментация смысла), тем длиннее текстовые конструкции и больше вероятность потери смысла в массе деталей, и тем больше Колмогоровская сложность алгоритма [4] процесса воспроизводства описываемого объекта.

По аналогии с фрактальным представлением береговой линии [5] описание объекта или процесса в документе будем представлять как ломаную линию, огибающую линию (идеальную) смысла. Соответственно, языковые конструкции будем рассматривать с точки зрения недостижимости эквивалентности образа (в данном случае текстового описания) оригиналу, а анализ выполнять не столько с точки зрения регулярности конструкций (что свойственно математическим фракталам), сколько в контексте приближения к истинному смыслу.

Для экспериментальных исследований и иллюстраций использовался массив статей по трём различным областям знания из научной электронной библиотеки Cyberleninka [6] и база данных диссертаций с семантическими поисковыми образами [7] в виде словосочетаний, соответствующих дескрипторам тезауруса (возможно, уточненным окружающими словами), а также словосочетаний, значимых для документа.

СИСТЕМНО-ФУНКЦИОНАЛЬНЫЙ ПОДХОД К КОЛИЧЕСТВЕННОЙ ОЦЕНКЕ РАЗМЕРНОСТИ СЛОВСОЧЕТАНИЯ И ПРЕДЛОЖЕНИЯ ЕСТЕСТВЕННОГО ЯЗЫКА

Язык, как средство коммуникаций, необходимо рассматривать не только как среду, обеспечивающую обмен сообщениями между субъектами в пространстве и во времени. Его другая, *когнитивная*, функция – это моделирование реальности, устанавливающее связь между пространствами образов и оригиналов.

Исходя из принципов систематизации, по признаку «статичность/динамичность» на этом «модельном» уровне (понятийно-знаковом) можно выделить две сущности (категории): объекты и процессы. Объекты, представляющие сравнительно устойчивые (статичные) образования, определяются, прежде всего, внутренними свойствами – проявлениями структуры. Процессы, представляющие динамику, определяются внешними обстоятельствами¹ – проявлениями взаимодействий объектов и процессов.

Основными единицами смысловыражающих конструкций будем считать словосочетания и предложения ЕЯ. Словосочетание будем рассматривать как устойчивое (по крайней мере в конкретном тексте) сочетание слов для обозначения *целостного* и *статичного* объекта (обычно, ассоциируемого с поняти-

ем), рассматриваемого как самостоятельная «конструкционная» единица в дальнейших построениях смысла. Предложение (а также абзац, глава, документ и т.п.) – это процессно-ориентированная конструкция для построения *композиционных* фрагментов смысла, позволяющая передавать *динамику процесса формирования* нового смысла.

Язык, как знаковая система, устанавливает соответствие знаков реальным или абстрактным предметам (объектам, процессам, свойствам). При этом, независимо от способа обозначения (номинативного или функционального), знаки наделяются признаками (характеристическими, отличительными, валентными), характеризующими соответствующие предметы с точки зрения их идентификации.

Слова являются минимальными единицами, обозначающими понятия, однако большинство понятий обозначается сочетаниями слов. Необходимо отметить, что смысловые значения сложных конструкций получаются непрямым суммированием смысловых значений слов. Смысл слова, добавляемого в конструкцию, «трансформируется», происходит связь его свойств (признаков) со свойствами (также признаками) других слов словосочетания или предложения. Это и является проявлением эмерджентности – системного свойства языка: словосочетание и предложение, как конструкция, составленная из слов, каждое из которых обозначает какой-либо объект/процесс/свойство, в совокупности дает новое качество и, соответственно, признак. Для иллюстрации этого свойства далее предлагается «наивная» векторная модель, представляющая значение смысловыражающей конструкции в двумерном пространстве с осью «реальных» – состоявшихся (т.е. уже имеющих однозначное значение) признаков и осью «мнимых» – синтетических признаков, отражающих новизну, образованную собственно сочетанием.

Таким образом семантика слов устроена так, что для того, чтобы любая другая часть речи приобрела свое значение, ей необходимо вступить в связь со смыслом другого слова («свое» строится на «чужом»), которое бы семантически верно отражало данное касательство [8].

В частности, словосочетания образуются при добавлении к главному слову, являющемуся существительным, зависимых слов, относящихся к разным частям речи. При добавлении зависимых слов конструкция уточняется. Причём уточняться может как сам объект (выраженный главным словом), так и его компоненты (зависимые слова). Именно поэтому при добавлении слова появляется признак, отражающий изменение образа главного объекта.

В основе технологии формирования словосочетания лежат определённые лингвистические шаблоны [9], например:

существительное + существительное в родительном падеже;

прилагательное + существительное;

существительное + прилагательное в родительном падеже + существительное в родительном падеже и др.

¹ Обстоятельства также фиксируются в виде свойств.

Для сравнительной количественной характеристики смысловыражающих конструкций примем следующие предположения о наборах признаков, свойственных разным частям речи вследствие их «функциональных» ролей.

Качественным прилагательным припишем один признак (сравнительная характеристика объекта или процесса).

Притяжательным прилагательным припишем три признака (признак существа объекта/процесса/свойства; признак валентности, обеспечивающий за счет связывания новое качество; признак – указание на конкретный экземпляр).

Относительным прилагательным припишем четыре признака (признак существа объекта/процесса/свойства; признак валентности; признак, указывающий на класс, и признак, отражающий системные свойства класса – возможность появления нового свойства).

Причастию припишем два признака («функциональный» признак, полученный от глагола; признак валентности, характеризующий передачу свойств действия субъекту или объекту).

Существительным в косвенных падежах припишем два признака (признак свойства инфинитива, от которого образована словоформа; признак, характеризующий условия использования объекта).

Наречиям припишем один признак (признак условий процесса).

Глаголу припишем два признака («функциональный» признак, характеризующий собственно действие; признак валентности, отражающий способность к связыванию).

Для существительного, являющегося главным словом, число признаков примем равным пяти – наибольшему количеству по отношению к перечисленным выше другим частям речи, поскольку сам объект имеет основное значение.

ФРАКТАЛЬНОСТЬ РАЗМЕРНОСТЬ И ВЕКТОРНАЯ МОДЕЛЬ ТЕКСТА

Проведём аналогию с предложенной Мандельбротом [10] аппроксимацией степени «убегания» длины береговой линии L в зависимости от шага циркуля δ :

$$L(\delta) = \alpha \delta^{1-D}.$$

В рассматриваемом контексте размерность D может служить характеристикой обобщения или детальности с точки зрения представления смысла различными текстовыми конструкциями.

Фрактальная знаковая структура – это бесконечное производство знаков, каждый раз происходящее в новом масштабе [5]. В качестве мерительного инструмента масштаба выбираются признаки, а в качестве предмета измерения – смысл конструкции. Меры D текстовых конструкций различны и «степень извилистости» или «степень убегания смысла» зависит от влияния признаков на изменение смысла объекта.

Основываясь на формуле фрактальной размерности, так же, как и для береговой линии, кривой Коха,

треугольника Серпинского, можно вводить меры, характеризующие фрагментарность смысла для словосочетаний и предложений.

Рассмотрим геометрическую иллюстрацию векторной модели изменения смысла главного слова (исходного объекта), условно отождествляя слова с наборами признаков (см. выше), привносящих новую составляющую в модифицируемую конструкцию.

Объединение признаков будем проводить по правилу сложения векторов. Причем, исходя из принципа независимости признаков слагаемых, будем считать (в пределе), что векторы перпендикулярны. Таким образом, признаки главной смысловыражающей единицы (а для словосочетания – главного слова), которая всегда имеет конкретный смысл, будем откладывать по оси состоявшихся признаков, а признаки добавляемых единиц – по перпендикулярной «мнимой» оси синтетических (новых) признаков. Для добавления следующей единицы (слова) ранее полученный вектор суммы будем считать исходным и переносить его по полученному направлению (осуществляя вращение координат) как по новой оси состоявшихся признаков.

Это достаточно наивная модель, однако она вполне отражает тот факт, что сложность идентификации объекта будет определяться не только количеством добавляемых признаков, но и направлением. Считая каждое слово самостоятельной единицей, при объединении слов получаем конструкцию, которая в значительной степени наследует смысл исходных слов, однако в совокупности обретает другое значение, что отражается координатой, включающей уже и «мнимую» компоненту, фиксирующую наличие новизны.

Размерность словосочетаний

По аналогии с математическим определением фрактальной размерности [10] и основываясь на том, что добавление следующего слова в пределе обеспечивает бесконечное уточнение главного, введем следующий показатель размерности (отражающий меру сложности) для словосочетаний:

$$D^s = - \frac{\ln N}{\ln \left(\frac{\varepsilon}{\sum_{i=1}^N \varepsilon_i} \right)}. \quad (1)$$

где N – количество слов; ε – количество признаков главного слова, обычно являющегося существительным; $\sum_{i=1}^N \varepsilon_i$ – суммарное количество признаков всех слов.

Эта зависимость отражает тот факт, что формирование нового смысла словосочетания определяется нелинейностью вклада последнего добавленного слова в смысловую составляющую конструкции.

Для графической иллюстрации динамики смысла, определяемой вкладом признаков добавляемых слов, использовались семантические индексы базы данных диссертаций. На основе формулы (1) были посчита-

ны размерности словосочетаний, а также построены векторы признаков, показывающие изменение смысла этих конструкций. Соответствующие иллюстрации приведены на рис. 1 и 2².

Можно заметить, что при увеличении размера конструкции увеличивается и её степень «детальности», размерность конструкции начинает уменьшаться, что соответствует потере смысла исходного объекта в массе деталей.

Помимо лингвистических шаблонов, словосочетание может быть представлено и как сочетание словосочетаний («сложение» словосочетаний).

В этом случае может быть принято следующее правило построения векторной модели: выделяются две и более конструкции словосочетаний, связанных между собой, например, отглагольным существительным. Связывающие слова присоединяются к словосочетанию, становясь его частью, а далее к смыслу уже состоявшейся конструкции присоединяется смысл второй конструкции, образуя своим объединением новый смысл всего сложного словосочетания (рис. 3, 4). В случае, когда отсутствует связывающее слово (рис. 5), объединяются смыслы двух и более уже состоявшихся конструкций.

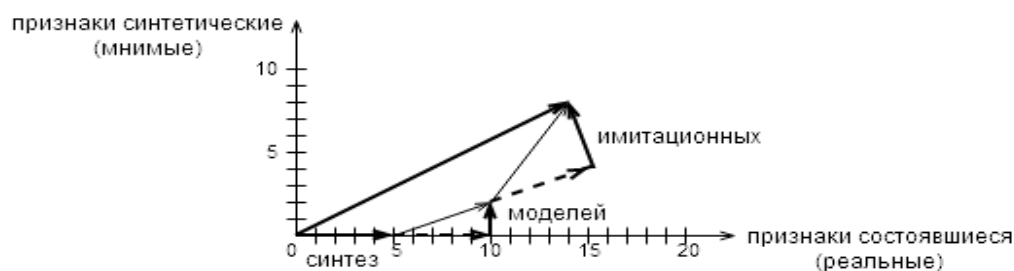


Рис. 1. Вектор признаков для словосочетания *синтез имитационных моделей* ($D=1,39$)

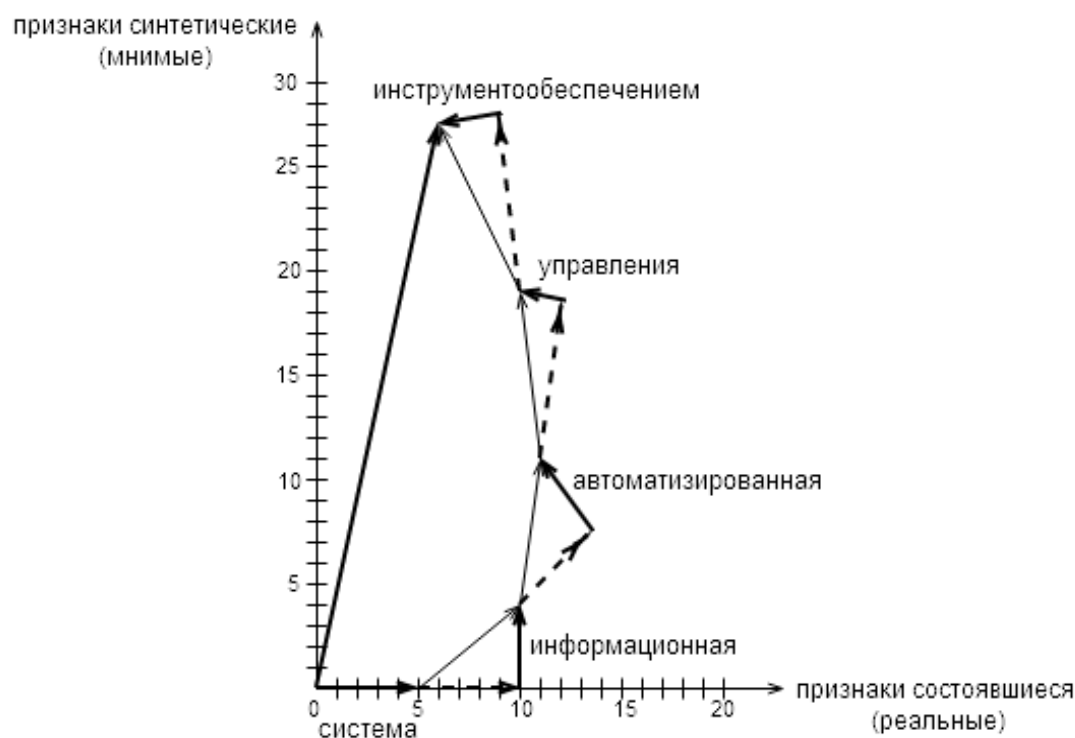


Рис. 2. Вектор признаков для словосочетания *автоматизированная информационная система управления инструментами* ($D=1,32$)

² Далее по тексту на рис. 1-8 жирной линией выделяются вектор признаков главной смысловыражающей единицы и перпендикулярно ему — вектор признаков добавляемого к ней слова. Обозначаемый тонкой линией вектор результирующего смысла переносится на ось состоявшихся признаков и обозначается пунктиром.

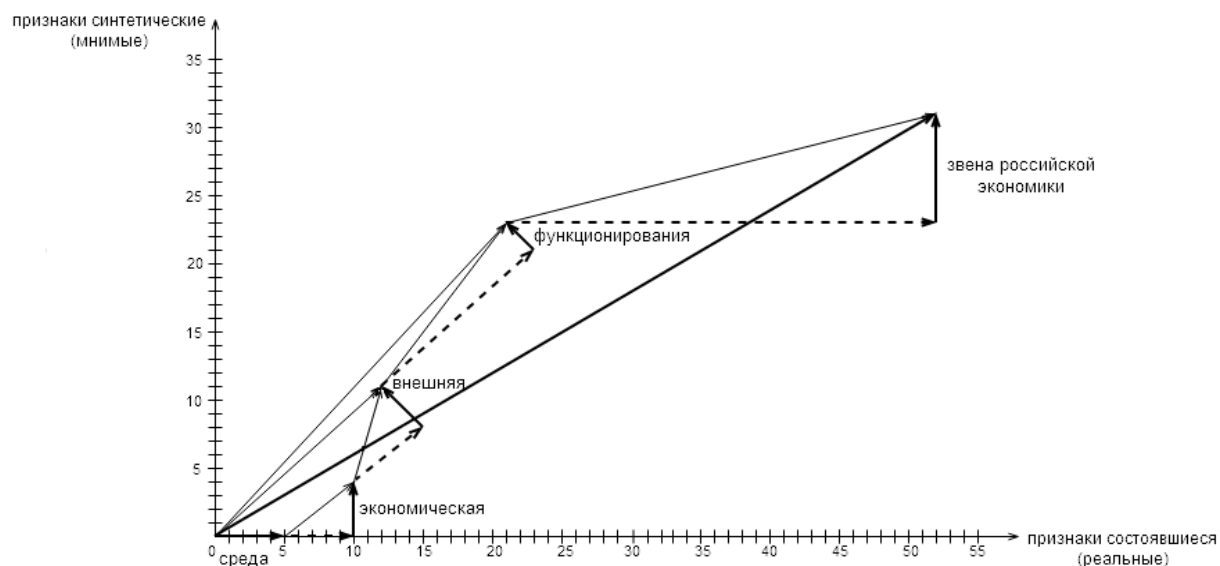


Рис. 3. Вектор признаков для сложного словосочетания *внешняя экономическая среда функционирования звена российской экономики*

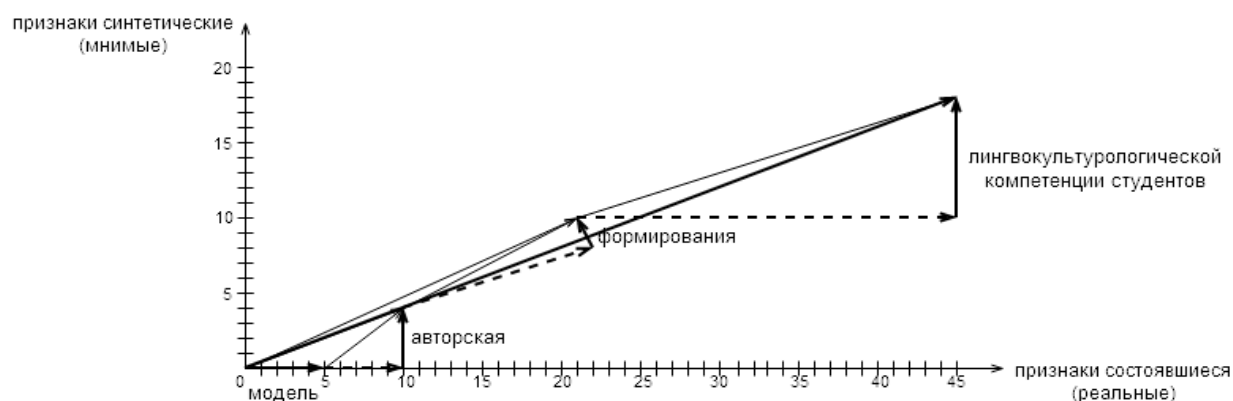


Рис. 4. Вектор признаков для сложного словосочетания *авторская модель формирования лингвокультурологической компетентности студентов*

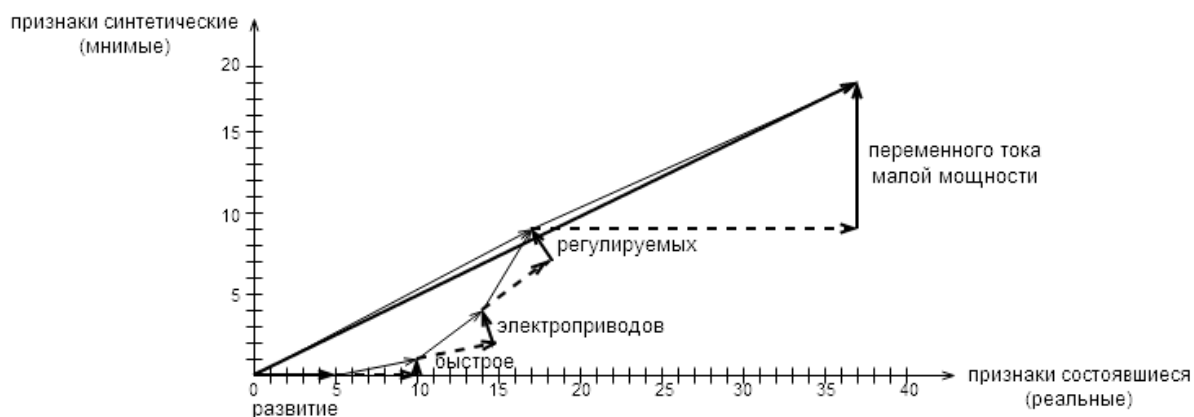


Рис. 5. Вектор признаков для сложного словосочетания *быстрое развитие регулируемых электроприводов переменного тока малой мощности*

Размерность предложений

С учетом выделенных словосочетаний размерность предложения определим следующим образом³:

$$D^P = - \frac{\ln M}{\ln \left(\frac{\sum_{p=1}^2 \varepsilon_p}{\sum_{i=1}^{M-N} \varepsilon_i + \sum_{k=0}^L D_k^S} \right)} \quad (2)$$

где M – количество слов в предложении; $\sum_{p=1}^2 \varepsilon_p$ – сумма признаков главных слов: ε_1 – количество признаков подлежащего, обычно являющегося существительным, ε_2 – количество признаков сказуемого, представленного чаще всего в виде глагола (знака

процесса); $\sum_{i=1}^{M-N} \varepsilon_i$ – суммарное количество признаков слов без учёта главных слов и слов из выделенных словосочетаний; $\sum_{k=0}^L D_k^S$ – суммарное значение размерностей словосочетаний, вычисляемых по формуле (1); L – количество словосочетаний в предложении.

Для сравнения были посчитаны размерности предложений научных статей технической, политической и гуманитарной областей, а также построены векторы признаков конструкций, отдельные примеры которых приведены на рис. 6 – 8.

Рассчитанные усредненные показатели размерностей имеют следующие значения: для политических текстов – $\langle D^P \rangle = 1,75$, для гуманитарных текстов – $\langle D^P \rangle = 1,82$, для технических – $\langle D^P \rangle = 1,88$.

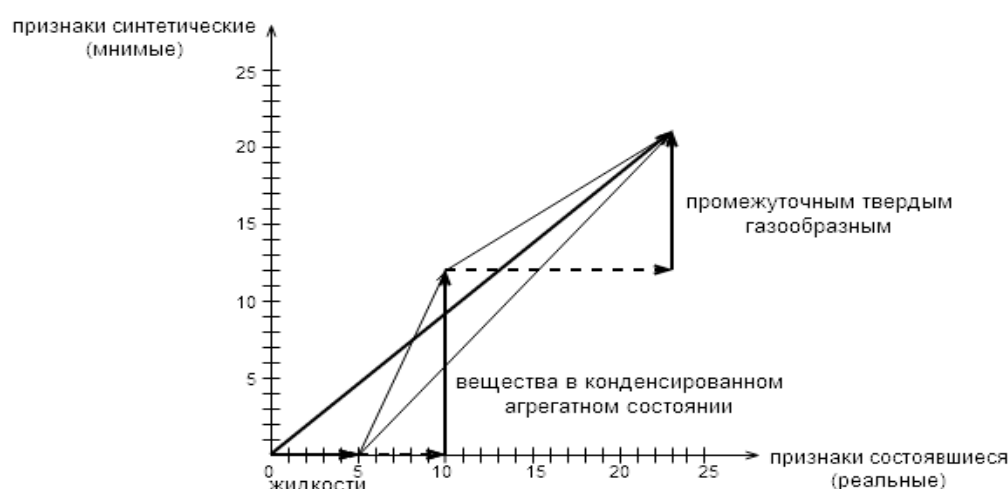


Рис. 6. Вектор признаков для предложения *Жидкости представляют собой вещества в конденсированном агрегатном состоянии, промежуточном между твердым и газообразным* ($D=1,87$)

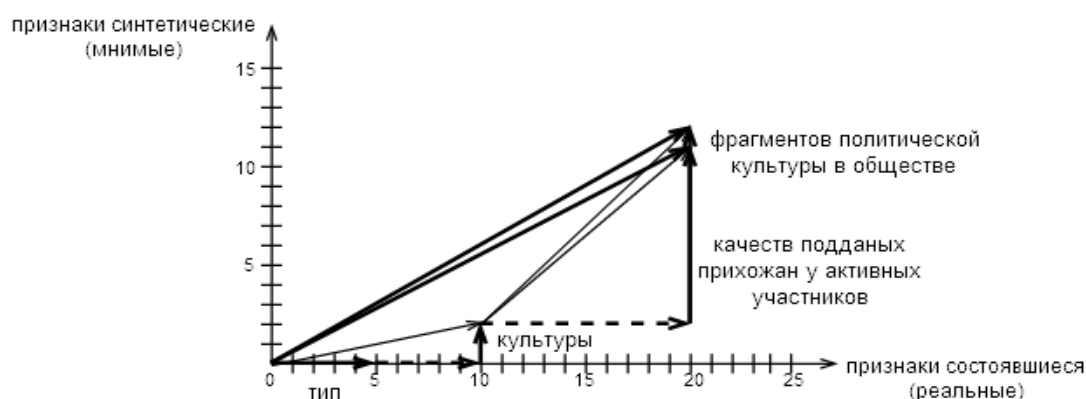


Рис. 7. Вектор признаков для предложения *Этот тип культуры предполагает, во-первых, наличие трех фрагментов политической культуры в обществе, во-вторых, наличие качеств подданных и «прихожан» даже у активных участников* ($D=1,72$)

³ Для сложного предложения мера сложности считается как сумма составляющих его мер.

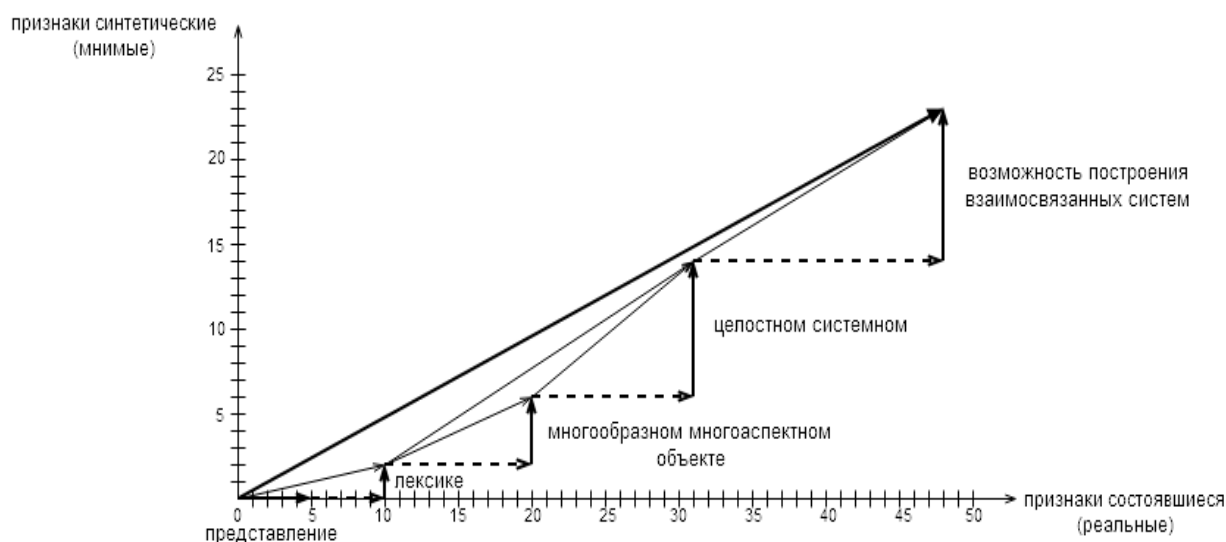


Рис. 8. Вектор признаков для предложения *Представление о лексике как о многообразном, многоаспектном и в то же время целостном системном объекте объясняет возможность построения взаимосвязанных её подсистем* ($D=1,79$)

Проведенные расчеты показывают, что минимальную размерность имеют предложения в политическом тексте. Это можно объяснить тем, что политические тексты отличаются наибольшей детальностью, обеспечивающей большую комбинаторность смыслов. Гуманитарные тексты обладают меньшей детальностью, поскольку в них используется специфическая лексика. Самая маленькая «детальность» и наибольшая размерность предложений у технических текстов. Это может быть объяснено тем, что в таких документах информация обычно представлена без лишних подробностей.

Отметим также, что добавление оборотов ведёт к росту синтетических признаков, а добавление однородных, уточняющих членов предложения приводит к росту состоявшихся признаков. У гуманитарных текстов наблюдается рост состоявшихся признаков, а у технических и политических — рост синтетических.

ЗАКЛЮЧЕНИЕ

Авторы настоящей статьи сознают, что предложенные решения не являются строго обоснованными, однако полученные результаты позволяют использовать системный подход для формализации понятия *сложность* текста и языка, которые рассматриваются как средство отображения (моделирования) действительности. Здесь понятие *размерности* может быть использовано как *интегральный* показатель детальности представления, позволяя на практике приводить разнообразие языка в соответствие с разнообразием предметной области, а также некоторым образом характеризовать сложность и разнообразие самой предметной области (точнее, её образа, представленного текстами). В частности, как показали примеры определения размерностей текстов из разных областей, максимальная «детальность» процесса

наблюдается у политических текстов, что объясняет минимальное значение показателя размерности.

Представленная векторная модель при всей своей упрощенности, тем не менее, достаточно очевидно отражает зависимость «успешности» (с точки зрения адекватности исходному смыслу) восприятия (интерпретации) сложных словарных конструкций: чем больше слов в словосочетании (более точно, детально представлен смысл), тем выше степень комбинативности и, соответственно, больше вероятность иной (по отношению к исходному смыслу) интерпретации текста. Это, в частности, подтверждается практикой нормализации словосочетаний, автоматически выделяемых из текста с помощью лингвистических процессоров: ошибки ложной координации слов растут в зависимости от количества слов в словосочетании. Такой подход может быть полезен для определения эффективной длины словосочетаний, используемых в качестве поисковых индексов: на сегодня нет ясности, что лучше — один термин, включающий более одного понятия (каждое из которых — словосочетание), или группа терминов, каждый из которых соответствует одному понятию.

СПИСОК ЛИТЕРАТУРЫ

1. Мильруд Р.П. В поиске имплицитного порядка языка как хаотической системы // Язык и культура. — 2009. — №4 (8). — С.60-65.
2. Филатова Е.В. Члены предложения и синтагма как реальная единица порождения и восприятия речи // Актуальные проблемы славянской филологии. — 2011. — №24. — С. 501-507.
3. Домброван Т.И. Фрактальность как фундаментальное свойство языковой материи // Записки з романо-германської філології. — 2013. — №30. — С. 61-68.

4. Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. – 1965. – Том 1. – С. 3-11.
5. Тарасенко В.В. Фрактальная логика. – М.: Книжный дом «ЛИБРОКОМ», 2009. – 120 с.
6. Научная электронная библиотека cyberleninka (КиберЛенинка), построенная на парадигме открытой науки (Open Science). – 2011. – URL: <http://cyberleninka.ru/article> (дата обращения: 12.06.2016).
7. Документальная информационно-аналитическая система xIRBIS. Свидетельство о гос. Регистрации программ для ЭВМ №2014619640 от 17.09.2014.
8. Лабикова Р.Н. Особенность структурно-семантической организации словообразовательных значений отадективных прилагательных русского языка // Актуальные проблемы гуманитарных и естественных наук. – 2014. – №4 (1). – С.1-13.
9. Величко В., Волошин П., Свитла С. Автоматизированное создание тезауруса терминов предметной области для локальных поисковых систем // “Knowledge–Dialogue–Solution” International Book Series “Information Science & Computing”. – 2009. – №15. – С. 24–31.
10. Мадельброт Б.Б. Фрактальная геометрия природы. – Ижевск: Институт компьютерных исследований, 2010. – 676 с.

Материал поступил в редакцию 25.07.16.

Сведения об авторах

ГОЛИЦЫНА Ольга Леонидовна – кандидат технических наук, доцент кафедры Системного анализа Национального исследовательского ядерного университета «МИФИ»
e-mail: OLGolitsina@YANDEX.RU

МАКСИМОВ Николай Вениаминович – профессор, доктор технических наук, профессор кафедры Системного анализа Национального исследовательского ядерного университета «МИФИ»
e-mail: NV-MAKS@YANDEX.RU

ЯКОВЕНКО Евгения Игоревна – студент Национального исследовательского ядерного университета «МИФИ»
e-mail: janeyakov@yandex.ru

Индексирование имен авторов испано- и португалоязычных стран в зарубежных журналах, а также в РЖ и БД ВИНТИ РАН

Имя лица (автора) является одним из терминов индексирования. Под именем подразумеваются фамилия, инициалы или личное имя (имена). Большие трудности возникают при индексировании (представлении) имен авторов из испано- и португалоязычных стран по правилам, позволяющим формировать ссылки на документы в различных источниках и адекватные поисковые запросы.

Ключевые слова: индексирование, имя автора, испано- и португалоговорящие страны

ВВЕДЕНИЕ

Имя лица (автора) является одним из терминов индексирования, выраженным одной или несколькими связанным лексическим единицам, представленным в поисковом образе документа и оформленным по правилам определенного информационно-поискового языка. Под именем лица подразумеваются фамилия, инициалы или личное имя (имена). К имени лица могут быть добавлены также и идентифицирующие признаки – ученая степень, должность, титул, а также слова «сын», «старший», «младший», цифры II, III и т.д., а также формальные, тематические, географические, хронологические подзаголовки. Это особенно важно в случае имени и фамилии, в которых возможны разночтения из-за языковой специфики и традиций в их представлении. В настоящей работе описываются особенности индексирования имен авторов из испано- и португалоязычных стран.

Журналы в англоязычных странах так же, как и журналы, частично или полностью издаваемые на английском языке в других странах, ежегодно получают все большее количество статей от неанглоязычных авторов. Было подсчитано [1], что количество статей авторов из Латинской Америки, отраженных в ISI Web of Knowledge, увеличилось вдвое с 1981 по 2000 гг. За правильное индексирование имен авторов несут ответственность издатели журналов, считает Билл Блэк [1]. Кроме выражения вежливости по отношению к автору это дает возможность обеспечить правильную ссылку на статью.

Большие трудности ожидают тех, кто встречает имена и фамилии авторов из испано- и португалоязычных стран. В испанском и португальском языках используются шаблоны построения имени, отличные от английского языка. Более того, правила испанского языка при формировании комплекса «имя+фамилия» отличаются от правил португальского языка. И испанские имена гораздо сложнее индексировать, чем португальские.

КЛАССИЧЕСКИЕ ПРАВИЛА ПРЕДСТАВЛЕНИЯ ИМЕН АВТОРОВ

Существуют стилистические руководства (например, [2, 3]) по «классическим» правилам представления испанских и португальских имен. Так, в именах испанского происхождения часто используются три составляющих: имя и две фамилии, первая из них – фамилия отца, а вторая – фамилия матери. Почти все без исключения такие «составные» имена должны быть проиндексированы с ориентацией на имя отца. Женщина, выходя замуж, теряет (убирает) фамилию матери, но оставляет фамилию отца. К этой фамилии добавляют/или не добавляют предлог *de* (аналог английского – *of*) перед фамилией отца мужа. Например, если María Esquivel López выходит замуж за Juan Sánchez Mendoza, то она теряет фамилию матери (López), оставляет фамилию отца (Esquivel), добавляет “*de Sánchez*” и теперь она – María Esquivel de Sánchez (возможен вариант без *de*). Оба – и мужчина, и женщина могут использовать союз *y* (“и”) при соединении фамилии отца и матери, и это будет выглядеть как Juan Sánchez y Mendoza, а не Juan Sánchez Mendoza. Иногда употребляется дефис для объединения двух имен, и это выглядит как Juan Sánchez-Mendoza или María Esquivel-Sánchez. Некоторые мужчины и женщины используют инициалы при обозначении фамилии матери, т. е. Juan Sánchez M. Португальцы также используют сложные имена, но в порядке, обратном испанскому, т. е. фамилия матери стоит на первом месте, а фамилия отца – на втором. Женщины варьируют написание своего имени после замужества, обычно опускают фамилию матери, оставляя фамилию своего отца и добавляя фамилию отца мужа. В любом из этих случаев при индексировании португальских имен следует использовать последнее слово в последовательности «имя + фамилия» автора, т. е. точно так же, как в английском языке. Но и тут бывают отклонения от правил. Дело в

том, к португальской фамилии могут быть добавлены идентифицирующие признаки: Junior – «младший» (в английском), Filho – «сын» (в португальском), Neto – «внук» (в португальском), II, III – «второй», «третий» (в английском) и т.д. И незнание того, что Filho (Neto) – вовсе не фамилия, а добавление к имени отца, приводит к тому, что при индексировании такого имени, Filho и Neto, согласно правилам использования последней фамилии как основной, трактуются как фамилии, что неизбежно порождает ошибки в ссылках, и невозможность найти место работы автора, чье имя было индексировано неверно. Как испанцы, так и португальцы используют сложные первые имена, например, María Cristina или José Fernando в испанском или Maria Tereza или Lúcia Helena в португальском. Составные имена могут также содержать предлоги и артикли, например, María de Jesús или María de las Mercedes в испанском и Maria das Dolores или Isabel da Conceição в португальском. Но при индексировании длинных имен в португальском языке не возникает проблем, потому что в этом случае используется фамилия отца, которая является последним словом в комплексе «имя+фамилия», так же, как и в английском языке (за исключением фамилий, заканчивающихся на Filho или Neto). Для людей же, не знакомых с испанскими именами, использование сложных имен представляет трудность: является ли слово, встречающееся в середине комплекса «имя+фамилия», второй частью имени или фамилией отца [4].

На протяжении трехлетнего периода, с января 2000 г. по декабрь 2002 г., редакция международного журнала "Revista Panamericana de Salud Pública/Pan American Journal of Public Health" [1] отслеживала поступающие англоязычные статьи. Это издание помещает публикации на английском, испанском и португальском языках, каждая статья появляется на одном из трех языков. Большая часть авторов журнала – из западного полушария, меньшая – из других стран мира, таких как Испания. Авторы поощряют писать статьи на своем родном языке, и, следовательно, языке, в котором они наиболее сильны. Но авторы могут представить материал и на другом языке. Например, автор из Бразилии может написать статью на английском или испанском языке вместо родного португальского. Все авторы англоязычных статей, которые в качестве аффилированной указали организации на испанском или португальском языке, были отброшены. К этой группе были добавлены авторы с испанскими или португальскими именами, аффилированные с международными организациями, базирующимися в англоязычных странах, для которых испанский или португальский является одним из официальных языков. Были исключены авторы с испанскими и португальскими именами, аффилированные с другими типами организаций в англоговорящих странах, например с университетами. Такие авторы могли родиться в англоязычной стране или проработать какое-то время в англо-ориентированной организации. Предполагалось, что они скорее будут следовать англоязычным стандартам написания имени, т. е. их имена будут индексированы так же, как и стандартные английские имена. Данное издание было

выбрано в [1] из-за уверенности в том, что среди авторов, пишущих для этого журнала на английском языке, окажется большое количество испано- или португалоязычных авторов, пишущих также для англоязычных журналов США или других стран. Некоторые авторы этого журнала получили ученые степени в англоязычных странах, пишут хорошо на английском и могут быть знакомы с различиями в представлении английских, испанских и португальских имен. Те же, кто никогда не жил вне своих родных стран, имеют ограниченные навыки в написании английских текстов, и поэтому платят профессиональным переводчикам для перевода своих статей на английский, и могут не знать правил представления имен в английском языке. Более того, хотя эти авторы постоянно публикуются на испанском или португальском языке в своих родных странах или где бы то ни было, появление публикации в этом издании может быть их первым опытом в написании англоязычной статьи для международного журнала.

«Классическим» испанским и португальским правилам в представлении имен следуют, естественно, не все авторы. Фактически, к большому облегчению издателей журнала "Revista Panamericana de Salud Pública/Pan American Journal of Public Health", авторы, чьи имена рассматривались в [1], доставили минимум хлопот для англоязычных редакторов. Например, из 90 выбранных испанских имен 41 (45,6%) состояло всего из одного имени и одной фамилии (например, Susana Vázquez), 15 (16,7%) – из имени, инициала, фамилии (например, María G. Guzmán). Анализ испанских имен из испаноязычных статей, опубликованных в 2000 г., продемонстрировал похожую картину: 54,4% и 10,7% соответственно. Остальные 6% (6,7%) представляют собой одно имя и составную (из двух частей) фамилию, части которой разделены дефисом (например, Héctor Díaz-Torres). Все эти три категории могут быть легко подвержены индексированию, следуя обычной практике представления англоязычных имен. Реально могут возникнуть проблемы с теми именами в испанском языке, в которых используются три или более слов без дефисов, или имеющих союзы, предлоги или артикли.

ИНДЕКСИРОВАНИЕ ИМЕН АВТОРОВ

При индексировании имен авторов важно уметь: 1) отличать испанский язык от португальского и 2) читать «элементы» испанских или португальских имен – знать, как обходиться с предлогами и артиклями, а также устанавливать различия между именами и фамилиями. При различении испанского и португальского имени наилучшим индикатором может быть страна автора. Испания и Португалия колонизировали большую часть Нового Света и принесли с собой свои языки. За исключением Бразилии, где говорят на португальском, все большие страны Центральной и Южной Америки используют испанский язык. Другая возможность научиться различать испанский и португальский языки – рассмотреть ударения и другие диакритические символы. Например, в испанском языке есть ñ (n с тильдой, как в фамилии Núñez), но португальский этого не имеет. В португальском есть ç – седиля (как в фамилии Gonçalves),

но испанском нет. Оба языка имеют ударения на гласных, но испанский применяет лишь одну форму, а португальский – несколько. Оба языка имеют ударение «акут», т. е. *é* в José – имени, встречающемся и в португальском, и в испанском языках. Португальский также применяет циркумфлекс, как *ô* в Antônio, и тильду над гласным для обозначения его назальной формы, как в João – португальской форме имени John. Полный перечень диакритических символов, используемых в португальском и испанском языках, дан в “The Chicago Manual of Style” [5].

Имея небольшую практику, можно, быстро взглянув на имя, написание и диакритические символы, точно установить, испанский это язык или португальский [6–8]. Например, любой человек быстро научится различать португальские имена João Gonçalves и Antônio Nunes и испанские – Juan González и Antonio Núñez. Но при рассмотрении испанского имени, имеющем в своей структуре более двух имен или более двух фамилий, как можно понять, что есть что? Например, как вы “прочитаете” имя María Celeste Gutiérrez? Как понять: Celeste – это часть имени или фамилия отца, а в Fernando Ruiz Meir: Ruiz – часть имени или фамилия отца, которые должны быть использованы для индексирования? Если автор указывает обе фамилии через дефис, это облегчает задачу. Например, в случае Juan Sánchez-Mendoza, автор будет проиндексирован составной фамилией Sánchez-Mendoza (сокращенная форма – Sánchez-Mendoza J.).

Рекомендации для индексирования имен основаны на правилах, приведенных в “Anglo-American Cataloguing Rules” (Англо-американские правила каталогизации) [9], которые используются большинством библиотек США и других англоговорящих стран для индексирования, а также для применения стандартных правил индексирования в испано- и португалоговорящих странах. Краткие формы, приведенные в Англо-американских правилах каталогизации, основаны на полных именах и соответствующих аббревиатурах согласно «Ванкуверскому стилю» (см. “Vancouver Reference Style Guide” [10]), используемому для ссылок в публикациях журналов биомедицинской тематики. В руководстве приводятся набор шаблонов, которые позволяют отличать имена от фамилий. Некоторые испанские имена имеют такое же написание, как и английские, например, Linda, Patricia, Irene, David, Daniel. Есть похожие, но использующие ударение (в испанском), например, Martín, или имеющие небольшие различия в написании, например, Eduardo и Sonia. Мужские имена в испанском языке чаще всего идентифицируются по окончанию *o* – Fernando, Rodrigo, Gonzalo. Женские имена чаще всего оканчиваются на *a* – María, Cristina, Anita.

Другой способ распознавания имен – использование их перечней. Существуют перечни испанских и португальских имен в печатном виде, а также в Интернете (таких ресурсов масса), например: Spanish baby names, Baby Name Guide, Portuguese Baby Names Behind the name. Portuguese Surnames, Iberian naming customs и пр. [11–16]. Есть одно полезное правило распознавания фамилий: многие из них образуются

прибавлением *-ez* к имени, как, например, Rodríguez из Rodrigo или González из Gonzalo. Приведенные примеры имен должны индексироваться так: María Celeste Gutiérrez по фамилии Gutiérrez (сокращение должно выглядеть как Gutiérrez M.C.), а Fernando Ruiz Meir – составной фамилией Ruiz Meir (в сокращенном виде Ruiz Meir F.).

Если испанское имя имеет союзы, предлоги и артикли, это может вызвать определенные затруднения. Среди самых общих (часто встречающихся) форм такого рода могут быть союз *y* (и); предлог *de* (аналог английского *of*); комбинации предлога и артикла *del*, *de la*, *de los*, и *de las*, а также артикли *Los* и *Las* (может встретиться написание *los* или *las*). Предлог *de* и связанные комбинации предлог-артикл (*del*, *de la*, *de los*, и *de las*) могут использоваться с фамилией отца и/или матери. Как и в других испанских именах, имя автора должно быть проиндексировано с использованием имени отца, например, José del Pozo García индексируется как Pozo (в сокращении – Pozo García J. del). В сокращенной форме оставляется полная форма предлога или комбинации предлог-артикл, но ставится после имени. Аналогичным образом, при наличии фамилии матери, использующей *de* или одну из комбинированных форм, имя автора все же должно быть проиндексировано именем отца. Например, Juan González del Moral должен быть проиндексирован как González (сокращенная форма выглядит как González del Moral J.). Если имя используется без предлога, то артикли *Los* и *Las* должны при фамилии остаться. Например, автор Manuel Antonio Las Heras должен быть проиндексирован как Las Heras (сокращенная форма должна выглядеть как Las Heras M.A.).

Португальские имена представляют собой меньше сложностей, чем испанские. Многие португальские авторы используют форму, очень похожую на английскую: одно имя, одна фамилия и, возможно, один – два инициала между ними. Если же автор указывает составное имя и составную фамилию, то последнее имя будет использовано для индексирования, как это обычно происходит в английском языке. Например, Neusa Maria Costa Alexandre будет проиндексировано с использованием фамилии отца Alexandre (сокращенная форма может выглядеть как Alexandre N.M.C.). При наличии предлогов и артиклей или их комбинаций могут возникнуть определенные трудности, в том случае, когда они образуют часть фамилии отца и имеют предлоги *de*, *da*, и *do* и комбинацию предлог – артикл *dos*. Например, должно ли имя Boaventura Antônio dos Santos быть проиндексировано как Santos или dos Santos? Если фамилия отца содержит формы *de*, *da*, *do*, или *dos*, то имя индексируется тем словом, которое за ними следует. Например, Boaventura Antônio dos Santos должно быть проиндексировано как Santos (в сокращении это будет Santos B.A. dos). С фамилией и предлогом варианты могут быть разные. Иногда союз *e* (и) объединяет два отдельных имени, а иногда образование «фамилия + и + фамилия» становится «фиксированным», как одна фамилия.

Один и тот же автор может использовать и не использовать союзы. При наличии союза «и» лучше всего выяснить у самого автора, что он предпочитает. Другой трудностью португальского языка является

присутствие таких форм, дополняющих фамилию, как сын (Filho) или внук (Neto) (в английском – это Jr – младший; Sr – старший; II, III – второй, третий и т.д.). Они должны быть частью фамилии, а не рассматриваться как последняя часть – как фамилия отца. Некоторые авторы проясняют этот момент, ставя дефис между фамилией и Filho или Neto как в Djalma S. Moreira-Filho, что будет проиндексировано как Moreira-Filho (сокращенная форма Moreira-Filho D.C.). Другие авторы, например, Alberto Pellegrini Filho и Antônio Ribeiro de Castro Neto, не использовали дефис. Их имена должны быть проиндексированы как Pellegrini Filho и Castro Neto (сокращенной формой будет Pellegrini Filho A. и Castro Neto A.R. de).

Испанские исследователи R. Ruiz-Perez, Delgado E. Lopez-Cozar и E. Jimnez-Contreras в [2] предлагают много полезных идей по поводу применения испанских имен. Авторы статей, считают они, должны поощрять к использованию одного написания имени-фамилии в течение своей карьеры и формы, близкой к английской модели: одно или два имени и одна фамилия. Предлоги, артикли, союзы должны опускаться. При использовании двух фамилий эти исследователи рекомендуют соединять их дефисом. Другой разумный совет: не употреблять в конце фамилии на месте фамилии матери инициалы (как в Carlos Suárez N.); лучше дать либо фамилию отца, либо обе фамилии через дефис. Некоторые из этих предложений применимы также и к португальским фамилиям. Более того, они предлагают автору использовать одну форму для комплекса «имя+фамилия» в его профессиональной деятельности. Авторы, использующие Filho или Neto, обязательно должны присоединять их дефисом. В зависимости от того, сколько журнал получает статей для печати из испано- и португалоязычных стран, редакторы должны включать правила оформления фамилий в методические рекомендации для авторов. Или при необходимости согласовывать с авторами правильность передачи их имен.

СВЯЗЬ ВАРИАТИВНОСТИ ПРЕДСТАВЛЕНИЯ ФАМИЛИЙ И КОЛИЧЕСТВА ОПУБЛИКОВАННЫХ СТАТЕЙ

Опытным путем в [2] определено, что существует непосредственная зависимость между вариантами представления имени и количеством опубликованных статей. С этим связана возможность с помощью одного поискового запроса найти все статьи автора, независимо от того, правильно или неправильно указано его имя в поисковом запросе. При этом были рассмотрены работы 172 авторов с Медицинского факультета Университета г. Гранады, проиндексированные в течение 10 лет в SCI (Science Citation Index), MEDLINE и IME (Indice Médico Español). Выяснилось, что авторы, опубликовавшие одну статью, имеют, естественно, один вариант написания имени. С ростом количества опубликованных статей растет также количество вариантов представления имени автора. Корреляция значительна: от 25% до 40% авторов с одной или двумя статьями публикуются более чем под одним вариантом имени. Более 50% авторов, имеющих от 5 до 10 публикаций, индексируются под более чем одним именем. Если у авторов 9 публикаций, то более 70% из них имеют несколько различных написаний имен, т. е. в такой ситуации с увеличением количества публикаций эффективность поиска статей конкретного автора резко падает.

В связи с упомянутым исследованием интересно будет рассмотреть пример поиска в БД Scopus статей автора Maciel Filho. Такой поиск дает не менее 25 вариантов представления имени (среди них – 7 ошибок в индексировании), собранных в 2-х профилях одного и того же автора: в одном случае обнаружилось 358, в другом – 52 документа. Статьи этого автора подобраны по тематике и названию аффилирующей организации. Но, тем не менее, варианты написания имени представлены в двух разных профилях автора, имеющих два разных идентификатора (рис. 1, 2).

Maciel Filho, Rubens Maciel lel	About Scopus Author Identifier View potential author matches
Universidade Estadual de Campinas, School of Chemical Engineering, Campinas, Brazil	Other name formats: Maciel Filho, Rubens
Author ID: 7003732915	Filho, R. M. Maciel Filho, R. Filho, Rubens Maciel Filho, Rubens Mac lel Filho, R. Maciel Maciel Filho, Rubens Maciel-Filho, Rubens Filho, Rubens M. FILHO Maceil Filho, R. Maciel-Filho, Ruben Filho, R. Mac lel Maciel Filho, R. Maciel-Filho, Rubens Filho Maciel Filho, Rubens Mac lel View Less

Рис. 1. Профиль 1. – Представление имени автора в БД Scopus

Maciel Filho, Rubens

Universidade Estadual de Campinas, School of
Chemical Engineering, Campinas, Brazil

Author ID: 56187460500

[About Scopus Author Identifier](#) | [View potential author matches](#)

Other name formats: Maciel Filho, Rubens

Maciel Filho

Maciel, Rubens

Maciel Filho, R.

Maciel-Filho, Rubens

Filho Maciel

Maciel

Maciel, Rubens

Maciel

Maciel Jr., Rubens

Maciel Filho

[View Less](#)

Рис. 2. Профиль 2. – Представление имени автора в БД Scopus

Большинство баз данных и служб индексирования содержат примеры статей, в которых были допущены ошибки при индексировании португальских фамилий, содержащих Filho или Neto в качестве составной части, стоящей на последнем месте. Пробный поиск в Базе данных PubMed (англоязычная текстовая база данных медицинских и биологических публикаций, созданная Национальным центром биотехнологической информации – NCBI) [17] дал следующие результаты: Filho встретилось в 1914 документах и Neto – в 2845.

Идентифицирующие признаки «сын» Filho или «внук» Neto должны всегда приводиться вместе с фамилией, входя в ее состав. Иначе нет никакой уверенности, что при поиске будут найдены все документы требуемого автора.

Для понимания, насколько серьезно подходят издатели к проблеме правильного индексирования имен, нами были выборочно просмотрены инструкции для авторов в 50 латиноамериканских журналах, находящихся в открытом доступе в электронной библиотеке SciELO – Scientific Electronic Library Online [18] и публикующих статьи на испанском, португальском и английском языках. Выяснилось, что единственной рекомендацией для авторов в этих инструкциях является представление имени и фамилии полностью. Это естественное пожелание, потому что такой подход дает возможность индексировать имена согласно рекомендуемым правилам.

Тем не менее, в самих изданиях даже в пределах одной статьи (а тем более внутри одного выпуска) появляются имена одних и тех же авторов в разных видах.

Ярким примером отсутствия в издании стиля в представлении блока «имя + фамилия», не говоря о том, что часто неясно, по правилам ли проиндексировано имя, является журнал “Lilloa” (Аргентина). Примеры даны по одному лишь выпуску – 2009 год,

т. 48, № 1-2, где приводятся списки авторов в разных статьях данного выпуска.

1) Andrada A.R.; M.E. Lozzia; M.E. Cristobal – у первого автора сначала идет фамилия, а затем инициалы (и имени, и фамилии), у последующих – в обратном порядке – сначала инициалы, потом фамилия;

2) Cristobal, Maria E., Maria E. Lozzia; Valeria de los Angeles Paez – у первого автора идет фамилия, затем имя и инициал (часть фамилии), у второго – имя, инициал (часть имени, вероятно, поскольку это испанский автор), фамилия, *причем явно в первом и втором примерах это один и тот же человек*; у третьего автора имя и фамилия расписаны полностью. В другой статье этого же издания этот же автор называет себя Valeria de los A. Paez. И, очевидным образом, имя проиндексировано неверно, исходя из правил испанского языка.

В SciELO существует стандарт публикации статей и оформления библиографических ссылок в различных библиографических форматах [см. 19].

Мы выбрали для примера различные варианты представления статьи из журнала “Revista brasileira de meteorologia” (2016, Vol. 31, № 1):

1. Название статьи и перечень авторов в том виде, в котором они представлены в полном тексте в pdf-формате (см. рис. 3).

2. Название статьи в журнальной базе данных SciELO (авторы проиндексированы) (см. рис. 4).

3. Название статьи и окно работы со ссылками (авторы как в первом примере) (см. рис. 5).

4. Описание того, как следует цитировать документ с использованием некоторых библиографических форматов, что позволяет заранее избежать ошибок при цитировании статьи в других источниках (см. рис. 6).

Данная статья представлена в БД Scopus. И мы видим ошибки в индексировании имен: dos Santos, C.A.C. вместо Santos, C.A.C. dos и de Brito J.I.B. вместо Brito, J.I.B. de [ошибка источника] (см. рис. 7).

Tendências de Índices de Extremos Climáticos para o Estado do Amazonas e suas Relações com a TSM dos Oceanos Tropicais

Carlos Antonio Costa dos Santos, Maria Monalisa M. S. Melo, José Ivaldo Barbosa de Brito

Рис. 3.

SANTOS, Carlos Antonio Costa dos; MELO, Maria Monalisa M. S. and BRITO, José Ivaldo Barbosa de. Trends in Climate Extremes Indices for the State of Amazonas and its Relationship with SST of Tropical Oceans. *Rev. bras. meteorol.* [online]. 2016, vol.31, n.1, pp.1-10. ISSN 0102-7786. <http://dx.doi.org/10.1590/0102-778620130001>.

Рис. 4.

Revista Brasileira de Meteorologia

versão impressa ISSN 0102-7786

versão On-line ISSN 1982-4351

Rev. bras. meteorol. vol.31 no.1 São Paulo jan./mar. 2016

<http://dx.doi.org/10.1590/0102-778620130001>

ARTIGOS

Tendências de Índices de Extremos Climáticos para o Estado do Amazonas e suas Relações com a TSM dos Oceanos Tropicais

Trends in Climate Extremes Indices for the State of Amazonas and its Relationship with SST of Tropical Oceans

Carlos Antonio Costa dos Santos¹, Maria Monalisa M. S. Melo¹, José Ivaldo Barbosa de Brito¹

¹Unidade Acadêmica de Ciências Atmosféricas, Universidade Federal de Campina Grande, Campina Grande, PB, Brasil

Serviços Personalizados

Artigo

nova página do texto(beta)

Português (pdf)

Português (epdf)

Artigo em XML

Referências do artigo

Como citar este artigo

Curriculum ScienTI

Tradução automática

Enviar este artigo por email

Indicadores

Citado por SciELO

Acessos

Altmetric 0

Links relacionados

Compartilhar

Mais

Mais

Permalink

Рис. 5.

Como citar

Formatos Bibliográficos

Formato Documento Eletrônico(ISO)

SANTOS, Carlos Antonio Costa dos; MELO, Maria Monalisa M. S. e BRITO, José Ivaldo Barbosa de. Tendências de Índices de Extremos Climáticos para o Estado do Amazonas e suas Relações com a TSM dos Oceanos Tropicais. *Rev. bras. meteorol.* [online]. 2016, vol.31, n.1 [citado 2016-06-27], pp.1-10. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-77862016000100001&lng=pt&nrm=iso>. ISSN 0102-7786. <http://dx.doi.org/10.1590/0102-778620130001>.

Formato Documento Eletrônico(ABNT)

SANTOS, Carlos Antonio Costa dos; MELO, Maria Monalisa M. S.; BRITO, José Ivaldo Barbosa de. Tendências de Índices de Extremos Climáticos para o Estado do Amazonas e suas Relações com a TSM dos Oceanos Tropicais. *Rev. bras. meteorol.*, São Paulo, v. 31, n. 1, p. 1-10, mar. 2016. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-77862016000100001&lng=pt&nrm=iso>. acessos em 27 jun. 2016. <http://dx.doi.org/10.1590/0102-778620130001>.

Formato Documento Eletrônico(APA)

Santos, Carlos Antonio Costa dos, Melo, Maria Monalisa M. S., & Brito, José Ivaldo Barbosa de. (2016). Tendências de Índices de Extremos Climáticos para o Estado do Amazonas e suas Relações com a TSM dos Oceanos Tropicais. *Revista Brasileira de Meteorologia*, 31(1), 1-10. <https://dx.doi.org/10.1590/0102-778620130001>

Gerenciadores de referências

[Exportar para BibTex](#)

[Exportar para Reference Manager](#)

[Exportar para Pro Cite](#)

[Exportar para End Note](#)

[Exportar para Refworks](#)

Рис. 6.

Рис. 7.

ИНДЕКСИРОВАНИЕ ИСПАНО- И ПОРТУГАЛОЯЗЫЧНЫХ ИМЕН В ИЗДАНИЯХ ВИНТИ РАН

Всероссийский институт научной и технической информации, как любой информационный орган, при библиографической обработке документов сталкивается с аналогичными проблемами. В библиографическом описании документа, сопровождающего реферат в РЖ и БД ВИНТИ, основное внимание уделяется правильному представлению фамилии лица (автора), которая всегда приводится на языке оригинала на первом месте перед инициалами или личными именами. При наличии нескольких авторов указываются фамилии всех авторов, принимавших участие в создании документа. Описания составляются на основании ГОСТ 7.1-2003 «Библиографическая запись. Библиографическое описание Общие требования правила составления» [20], а также Российских правил каталогизации [21]. Мы пользуемся тем, что было переработано для российского потребителя. Рассматриваются простые и составные фамилии, а также фамилии с приставками, предлогами, артиклями и фамилии, имеющие идентифицирующие признаки (типа Junior (*Jr*), Senior (*Sr*), Filho).

Простые фамилии состоят из одного слова: Lopez, Diaz.

Составные фамилии включают два и более слов, соединенных между собой дефисами, частицами или без них: Escamilla-Rivera (дефис), Garcia De Torres (предлог *De*). Bezerra de Mello (предлог *de*). Если в фамилии есть предлоги (*de*, *da*), сочлененные предлоги (предлоги, объединенные с артиклем – *del*), то они приводятся перед фамилией: Da Silva, De Pra, Dos Santos, Del Val, Del Valle-Zermeno. Это правило отличается от описанного выше правила представления аналогичной конструкции José del Pozo García, которая индексируется по международным правилам как Pozo García J. del. Подобные фамилии приводятся полностью в последовательности: фамилия + имя: Gomez de la Cerna Ramon. Приставки, артикли, предлоги с артиклями, частицы в имени (фамилии) автора (авторов) сохраняются. Артикли могут приводиться как с прописной буквы, так и со строчной: Dos Santos, dos Santos. Слова Santa, San, Santos (различные производные слова «святой») всегда сохраняются перед фамилией лица, независимо от способа соединения с ней: Dos Santos Ferreira David, Santos Rodriguez F.J., San Cristoval E., Sant'Ana M.M.).

Идентифицирующий признак «сын» – Filho приводится вместе с фамилией, входя в ее состав, подоб-

но таким идентифицирующим признакам, как Junior (*Jr*), Senior (*Sr*), II, III и т.д. в скобках.

Filho недопустимо использовать в качестве самостоятельной фамилии. Так, написание W.L. Filho, представленное в иностранном журнале, ошибочно и требует правки. А, например, фамилия Junior I.D.S.Q. на деле должна была быть представлена как Souza Queiroz (*Jr*) Idalmir. Но не всегда существует возможность это расшифровать.

Другой особенностью представления испанских и португальских фамилий является соединение двух фамилий союзом, например: **Lopez y Garcia** (подчеркивается и выводится в авторский указатель как единая фамилия) или **Beraldo e Silva L.J., D'Albuquerque e Castro J.** (союзы *y*, *e* приводятся со строчной буквы).

Если оригинальная фамилия начинается с союза (например: **y Garcia**) – это также ошибка, требующая установления подлинной фамилии и написания ее вместо ошибочной.

Фамилии, состоящие из двух частей, приводятся как единая фамилия (в таком же виде выводятся в авторский указатель) без сокращения той или иной ее части до буквы. Например, фамилия **Hernandez Gomez** не может быть представлена в описании как **H. Gomez** или **Hernandez G.** (что часто встречается в изданиях и, соответственно, требует правки). Необходимо установить обе части фамилии в полном написании, независимо от того, как они представлены в документе (в авторские указатели все перечисленные случаи ошибок не должны попадать, не говоря уж о самих реферативных изданиях ВИНТИ).

ВЫВОДЫ

При индексировании сложных по конструкции испанских и португальских имен возникают большие трудности. Это проявляется чаще всего в англоязычных изданиях. Но довольно часто ошибки при индексировании имен собственных бывают характерны и для самих авторов – носителей языка. Ошибки возникают из-за незнания классических правил индексирования. Что касается ВИНТИ РАН, то жесткие правила, принятые ГОСТами, часто создают дополнительные трудности при оформлении документов, так как поиски в различных Интернет-ресурсах полного имени или частей имени автора при неправильном представлении в источнике комплекса «имя+фамилия» не всегда могут увенчаться успехом.

СПИСОК ЛИТЕРАТУРЫ

1. Bill Black. Indexing the Names of Authors from Spanish- and Portuguese-Speaking Countries // Science Editor. – Vol. 26, № 4. – URL: <http://www.councilscienceeditors.org/wp-content/uploads/v26n4p118-121.pdf> (дата обращения 25.04.16)
2. Ruiz-Perez R., Delgado Lopez-Cozar E., Jimnez-Contreras E. Spanish personal names variations in national and international biomedical databases: implications for information retrieval and bibliometric studies // J. Med. Libr. Assoc. – 2002. – Vol. 90, № 4. – P. 411-430. – URL: http://eprints.rclis.org/12869/1/Spanish_personal_name_variations_in_national_and_international_biomedical_databases_implications_for_information_retrieval_and_bibliometric_studies.pdf (дата обращения 15.05.16)
3. Naming Conventions of Spanish-Speaking Cultures. Understanding Spanish Family Names & Given Names. – URL: http://www.lrc.salemstate.edu/hispanics/other/Naming_Conventions_of_Spanish-Speaking_Cultures.htm (дата обращения 08.06.16)
4. Практическая транскрипция фамильно-именных групп. – Изд. 2-е, доп. – М.: Наука, 2006. – 526 с.
5. The Chicago Manual of Style. – URL: <http://www.libs.uga.edu/ref/chicago2009.pdf> . (дата обращения 08.05.16)
6. Гиляревский Р.С., Старостин Б.А. Иностранные имена и названия в русском тексте. – 3-е изд. – М.: Высшая школа, 1985. – 303 с.
7. Гиляревский Р.С., Старостин Б.А. Определитель языков мира по письменностям. – 3-е изд., перепеч. – М.: Наука, 1965. – 375 с.
8. Гиляревский Р.С., Гривнин В.С. LANGUAGES IDENTIFICATION GUIDE. – М.: Nauka, 1970. – 347 p.
9. Anglo-American Cataloguing Rules, 2nd Revision, AACR2. – URL: http://www.slideshare.net/Jamila-Jaber1/aacr2-14500111?from_action=save (дата обращения 25.05.16)
10. Vancouver Reference Style Guide. – URL: <http://openjournals.net/files/Ref/VANCOUVER%20Reference%20guide.pdf> (дата обращения 24.05.16)
11. “Spanish baby names”. – URL: <http://babynames.net/all/spanish> (дата обращения 15.05.2016)
12. Baby Name Guide. – URL: <http://www.babynamenameguide.com/categoryspanish.asp?strCat=Spanish> (дата обращения 15.05.2016)
13. Portuguese Baby Names. – URL: <http://www.babynamescube.com/portuguese-baby-names> (дата обращения 14.05.2016)
14. Behind the name. Portuguese Surnames. – URL: <http://surnames.behindthename.com/names/usage/portuguese> (дата обращения 25.05.16)
15. Iberian naming customs. – URL: http://www.tvwiki.tv/wiki/Iberian_naming_customs (дата обращения 04.06.16)
16. NationalMaster.com. – URL: <http://www.statemaster.com/encyclopedia/Spanish-naming-customs> (дата обращения 09.06.16)
17. PubMed. – URL: <http://www.ncbi.nlm.nih.gov/pubmed?term=Filho%5BAuthor%5D> (дата обращения 25.06.2016)
18. SciELO – Scientific Electronic Library Online – URL: <http://www.scielo.org/php/index.php> (дата обращения 16.05.16)
19. Хачко О.А., Глобачева Э.Я. Некоторые аспекты представления фамилий авторов испанского и португальского происхождения в зарубежных изданиях и базах данных, а также в РЖ и БД ВИНТИ // XVII Научно-практический семинар «Информационное обеспечение науки: новые технологии», Таруса, 24-28 июня, 2013. – С. 223-229. – URL: http://www.benran.ru/SEM/Sb_13/sbornik/223.pdf (дата обращения 15.04.16)
20. ГОСТ 7.1-2003. Библиографическая запись. Библиографическое описание. Общие требования и правила составления. – URL: <http://rusla.ru/rsba/provision/standarts/gost%207.1-2003.pdf> (дата обращения – 28.05.16)
21. Российские правила каталогизации / Рос. библиот. ассоц., Межрегион. ком. по каталогизации. – М., 2004. – 242 с. // Национальный информационно-библиотечный центр «ЛИБНЕТ». – URL: <http://www.nilc.ru/>

Материал поступил в редакцию 29.06.16.

Сведения об авторах

ХАЧКО Оксана Александровна – старший научный сотрудник Отделения обработки входного потока научно-технической литературы ВИНТИ РАН
e-mail: oxana@viniti.ru

ГЛОБАЧЕВА Эмма Яновна – главный специалист (исслед.), руководитель группы библиографической обработки научно-технической литературы ВИНТИ РАН
e-mail: oxana@viniti.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ в версии Windows Вы можете получить за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, Москва, ул. Усиевича, 20, ВИНТИ РАН

Коммерческое управление

Телефон/Факс: 8 (499) 155-45-25, 8 (499) 152-58-81

E-mail: contact@viniti.ru, sales@viniti.ru