МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 3** **Москва 2021**

ОБЩИЙ РАЗДЕЛ

УДК 316.32:33:004

А.П. Любимов, В.В. Черный

Эволюция глобализма: от компьтеризации – к электронной демократии и цифровой экономике знаний

Рассматриваются актуальные проблемы компьютеризации, электронной демократии и цифровой экономики знаний. Глобальные тренды в сфере информационного цифрового общества, электронной демократии и экономики знаний составляют повестку современного развития цивилизации. Новые возможности в условиях информационного общества показали, что виртульность способствует появлению необычной реальности, которая может рождать новые угрозы и кризисы. Обсуждаются сценарии, когда под влиянием виртуальности меняются экономика, наука, сфера инноваций, культуры и быта.

Ключевые слова: *электронная демократия, цифровая экономика, экономика знаний, национальная безопасность, глобализация, инновации, информационное общество, гражданское общество*

DOI: 10.36535/0548-0027-2021-03-1

ВВЕДЕНИЕ

Современные высокие технологии и системы все быстрее осваивают различные отрасли экономики, примеров можно привести много[1]. В начале популярной книги-эссе «Куда идет человечество? О тенденциях международных отношений в XXI веке» [1] Е.П. Бажанов и Н.Е. Бажанова ссылаются на китайскую мудрость, что «предсказывать сложно, особенно будущее», а далее они утверждают, что «будущее вообще непредсказуемо». Но затем и в тексте книги, и статьях [2, 3] их позиция смягчается, и авторы соглашаются с китайской мудростью. Действительно, главные события XX-XXI вв. подтверждают справедливость обоих пессимистических выводов. Как показывает исторический опыт, прогнозы путей развития человечества, как правило, не очень точны. И все-таки необходимо пытаться определить тенденции в развитии человечества, по крайней мере, в ближайшей перспективе, основываясь на анализе процессов, происходящих на планете, а затем рассмотреть возможности экономического развития современной России.

Мы живём в эпоху стремительного укрепления взаимосвязи и взаимозависимости государств. Эта тенденция получила название «глобализация» [3]. Интенсивно развивается фундаментальная наука, ускоряется научно-технологическая революция в модернизации средств транспорта, в информационных технологиях и коммуникациях, наноэлектронике и биомедицине, космических исследованиях и многих других областях. Процесс глобализации демократии изменяет картину современного мира [4-16]. Сегодня человечество переходит в состояние глобального информационного общества [10-14]. Уже отмечается новое явление: глобализация имеет циклический характер и сопровождается периодическими финансово-экономическими кризисами.

ЭЛЕКТРОННАЯ ДЕМОКРАТИЯ И ЦИФРОВАЯ ДИПЛОМАТИЯ

Компьютеризация общества и изобретение Интернета уже фактически привели к возникновению народной электронной демократии [7]. Об этом свидетельствуют многочисленные инструменты Интернета для выхода в глобальное информационное общество, доступные каждому человеку. Однако государства по-прежнему остаются главными игроками на международной арене[2]. Интернетизация мирового пространства и либерализация финансов делают границы государств фактически прозрачными[3]. Цифровая дипломатия становится инструментом «мягкой силы» для продвижения внешнеполитических взглядов и влияния на общественное мнение. Руководства стран постоянно призывают интенсивнее использовать новые технологии на разных платформах, включая социальные медиа, для разъяснения позиций государства. В то же время интеллектуальные ресурсы, инициативы гражданского общества и каждого индивида могут оказать влияние на тенденции экономического развития стран и всей системы международных отношений [14-18]. В условиях информационного общества открывшиеся возможности показали, что виртуальность способствует появлению необычной реальности, которая рождает новые угрозы и кризисы [5, 17] Для существования каждой страны и для выживания всего человечества возникает необходимость построения эффективной системы противодействия новым вызовам и угрозам[4].

Глобальные тренды в сфере формирования информационного общества и электронной демократии составляют повестку современного развития. Сегодня мировая пресса все больше пугает обывателя моральным упадком современных демократий и деградацией политсистем на целых континентах. С пессимизмом пишутся о хаосе и кризисе мировой экономики в самой сердцевине ее либерального ядра, о не оправдавшихся надеждах на выгодное продвижение по миру рыночных принципов. Почти с истерикой СМИ рассказывают о повышенной турбулентности в международных отношениях. С опаской – о нарастании социальных брожений и даже о возможном крахе мирового капитализма в том виде, в котором его нам представляют современные теоретики «новой цифровой экономики». На многочисленных научных конференциях обсуждают то неожиданное пришествие длинных кризисных «циклов Кондратьева», то пробуксовку зачатков «умной коммерции». Придумывают пугающие теории «убывающей экономики». Нагнетают извечную борьбу между приверженцами экономических теорий Кейнса и Фрид-

[1] Любимов А.П. От информации, информационных процессов и технологий до нанотехнологий. Интервью с Нобелевским лауреатом, депутатом Государственной Думы, академиком и вице-президент РАН Ж.И. Алфёровым // Представительная власть – XXI век. – 2009. – № 4. – С. 1-5; Любимов А.П. Формирование национальной концепции инновационной системы России (часть 1) // Представительная власть – XXI век. – 2011. – № 7-8. – С. 24-29; (часть 2). Там же. – 2012. – № 2-3. – С. 9-14; Любимов А.П. Основы электронной демократии в России // Актуальные вопросы экономики, управления и права: сборник научных трудов (Ежегодник). – М.: МГУКИ, 2013. – С. 5-11; Любимов А.П., Черный В.В. Антикризисная экономика и декриминализация банкротства // Представительная власть – XXI век. – 2019. – № 3. – С. 17-21; Черный В.В., Цыкало В.В., Аляев А.В. Россия и международная безопасность третьего тысячелетия // Обозреватель – Observer. – 2004. – № 5. – С. 27-35; Черный В.В. Безопасность России – в развитии инновационной экономики // Обозреватель-Observer. – 2009. – № 7. – С. 53-59; Черный В.В. Интеллектуальная революция и Россия // Стратегия России. –2009. – № 7. – С. 29-32; Черный В.В. Наука и глобальная демократия: роль России // Мир и политика. – 2009. – №11. – С. 101-109; Черный В.В. Глобализация демократии: гражданское общество и безопасность // Дипломатическая служба. – 2011. – № 5. – С. 12-20.

[2] Черный В.В. Глобализация демократии: интеллектуальная революция информационного общества и темная материя глобальных финансов // Представительная власть – XXI век. – 2012. – № 2-3. – С. 63-66; Там же. – 2012. – № 4. – С. 33-37; № 5-6. – С. 48-52; Черный В.В. Постоянный Форум ООН по правам коренных народов – 2019 // Представительная власть – XXI век. – 2019. – № 4. – С. 14-16.

[3] Черный В.В. Безопасность России – в развитии инновационной экономики // Обозреватель-Observer. – 2009. – № 7. – С. 53-59; Черный В.В. Интеллектуальная революция и Россия // Стратегия России. –2009. – № 7. – С. 29-32.

[4] Черный В.В. Выступление на 11th session of UN PFII, 07-15 May, 2012, Item 8.

мана, которые традиционно либо изощряют модели стимулирования спроса, либо предлагают новые, хитроумные способы опережающего производственного предложения и упреждающих инвестиций.

В сложных переплетениях предлагаемых теорий необходимо проанализировать и осознать, как все они объясняют причины современного этапа развития человеческого общества? И при этом выявить, что в предстоящем развитии будет главнее: информационное усиление рынка или его засорение, забивание его дыхательных пор и без того очевидным информационным загрязнением среды? Попробуем также разобраться, почему мировая экономика, едва успев взбодриться после того, как лопнули финансовые пузыри, тут же завязла в ухабах вновь образовавшегося кризисного бездорожья, почему главные локомотивы для прорывов глохнут на запасных путях и как тут исправить ситуацию для бескризисного движения? Для этого надо бы прочувствовать тренды в системе управления развивающейся демократии.

Сегодня понятно, что опыт внедрения либеральной экономики в России 1990-х годов с её реформами и приватизацией был экономически не эффективным. После приватизации предприятия становились планово убыточными, рабочие теряли работу, население нищало. Востребованную продукцию в Европу не пускали: там свои стандарты и высокая конкуренция. Страна начинала жить на денежные и товарные кредиты, которые в таких случаях давал Запад. Долги отдавали с процентами. Развитие страны становилось проблематичным, она жила по законам экономики периферийного капитализма. Опыт показывает, что либерализация экономики всегда происходит за счет бедных. Поскольку правительство отвечает за западные кредиты, а не за народ, то страна теряет государственную самостоятельность. Для выхода из такой ситуации надо выбирать зарубежных партнеров, сотрудничество с которыми будет соответствовать национальным интересам. А во внутренней экономике необходимо развивать предпринимательскую инициативу населения и государственно-частное партнерство.

В нашей концепции мы отталкиваемся от определенных вех в истории развития человечества, сыгравших судьбоносную роль в изменении жизни на планете. За точки отсчета мы берем аграрную и индустриальную революции, компьютерную революцию и интернет-революцию, связанную с интеллектуальной революцией информационного общества. Под термином «революция» мы понимаем кардинальное изменение технологического уклада, экономики и социальной жизни. Уже сегодня можно обозначить полюсы развития: виртуальность, с одной стороны, изменила науку и промышленность, а с другой – глубоко проникла в сферу политики и предопределила в ней даже гибридные войны. Мы наблюдаем лавинообразное «поумнение» – самостоятельное интеллектуальное развитие целых областей творческой деятельности и хай-тека.

Современное толкование понятия демократии зачастую хоть и считается несколько неопределенным, но все-таки оно опирается на три столпа: права человека; открытое мировое сообщество и открытый мировой рынок; гражданский контроль над вооружениями. Так считают ученые, по крайней мере, стран, более или менее успешно продвигающих демократию на международной арене. Процесс демократических преобразований в мире сложный, он широко обсуждается и заслуживает отдельного рассмотрения. Не останавливаясь на этом, отметим очень важный общий аспект демократических преобразований: сегодня необходимо учитывать быстрые изменения в мире, связанные с внедрением науки и высоких технологий в жизнедеятельность на планете. Поэтому новое определение понятия демократии должно включать эту роль еще одним, четвертым столпом – «наука и высокие технологии для международной безопасности и сотрудничества стран» [6]. Это позволит человечеству избежать пессимистических сценариев «конца истории», «столкновения цивилизаций» и других. В эпоху электронной демократии и экономики знаний развивающаяся мировая демократия не может выжить без спасительного голоса науки и высоких технологий.

ГЛОБАЛИЗАЦИЯ ИНДИВИДУАЛЬНЫХ ЧАСТНИКОВ

Не углубляясь в хитросплетения укрепления военной мощи стран и геополитических проблем, в нашем анализе важно выявить главные драйверы экономического роста жизни людей и понять их место в тенденциях современного планетарного развития. Неравенство экономического развития стран ведет к потере научно-технологического суверенитета, усложняет международные отношения, ослабляет международную и национальную безопасность, требует особого внимания к обеспечению военной безопасности, что ведет к увеличению затрат на укрепление обороноспособности.

Мы наблюдаем, как частная инициатива в сфере информационно-коммуникационных, микро- и нанотехнологий в Силиконовой долине выпестовала и оформила второе поколение уже частного освоения космоса типа *SpaceX* и *Virgin Galactic*. К 55-летию полета человека в космос российский бизнесмен Ю. Мильнер выделил 100 млн долларов *NASA AMES Research Center* на проект *Breakthrough Starshot* для изучения дальнего космоса и поиска жизни в районе звезды Альфа Центавра на основе идеи Ф. Цандера (1926 г.) о способе межзвездных путешествий. Сегодня разрабатывается проект посадки беспилотного Dragonfly на Титан, спутник Сатурна. 21 октября 2020 г. космический зонд НАСА *OSIRIS-REx* в автоматическом режиме завис на высоте 3,5 м над астероидом Бенну, выдвинул штангу и взял образцы грунта, а вернуться на Землю он должен в сентябре 2021 г. Чтобы понять гениальность этого эксперимента надо знать, что радиосигнал от Земли до астероида идет 18 минут. В 2021 г. в дополнение к уже работающему 30 лет орбитальному телескопу Хаббл НАСА запускает орбитальный телескоп Уэбб. Его зеркало диаметром 5,6 м робот будет собирать на орбите из отдельных сегментов. Вместе с 263 наземными телескопами и самым большим телескопом в Чили (с диаметром зеркала 39 метра) вся система будет не только изучать космос вглубь на 13 млрд световых лет, но и обеспечивать космическую безопасность Земли. В НИЦ «Курчатовский институт» реализуются концепции так называемых конвергентных НБИКС-технологий, объединяющих нано-, био-, информационные, когнитивные, да и социогуманитарные технологии, а также

и те гибридные отрасли хай-тека, которые были преобразованы виртуальным изменением мира.

Это означает приток интеллектуальных сил со всего мира, что заставляет говорить о глобализации демократии в научной сфере. Этот процесс проник в производство, усовершенствовав многие подходы в реализации естественных прав человека в свободном творческом научном поиске и в новой индивидуализированно-частной социальной организации труда.

Всё это позволило более широко реализовать индивидам их собственные идеи и дать свободу ученым для достижения состояния безбедного существования. Здесь полезными примерами являются особенно удачливые *Intel*, *Microsoft*, *Apple*, *SpaceX*, но есть и многие другие. Все они начинали с малых форм индивидуального частного предпринимательства, а сегодня стали гигантами бизнеса, определяющими уровень развития цивилизации на всей планете. Они появились как продукт развития культуры человеческой деятельности в области науки и высоких технологий, либерализации финансов, совершенствования креативных способностей для дальнейшего развития индивидуального творчества.

Инициатива новаторов, поддержанная государством, неоднократно приводила к новой технологической реальности, изменявшей жизнь людей на планете, к другому выстраиванию финансово-экономической системы, совершенствованию социальной организации труда и фактически – к новым технологическим укладам. Это были изобретатели паровой машины И. Ползунов и Дж. Уатт, создатели паровоза Г. Стефенсон и М. Черепанов, изобретатели радиосвязи А. Попов и Г. Маркони, создатели точечного транзистора У. Шокли, Г. Мур и Р. Нойс, основатель *Intell* Э. Гроув, основатель *Apple* С. Джобс, создатель *Microsoft* Б. Гейтс, основатель социальной сети *Facebook* М. Цукерберг, основатель *Google* С. Брин, пионер частной космонавтики И. Маск, создатель 400 компаний под крышей *Virgin Group* Р. Брэнсон, ученые-физики П. Капица, Л. Ландау, А. Абрикосов, В. Гинзбург, Н. Басов, А. Прохоров, М. Келдыш, А. Александров, Е. Велихов, Ж. Алферов, Ю. Гуляев и М. Ковальчук, математики В. Садовничий и Г. Перельман, ученые: химики, биологи, медики и другие.

Для государства здесь важным моментом становится умелое отслеживание тенденций опережающего развития и своевременное их внедрение в практику инноваций. Из опыта США мы видим хорошо организованный этап поддержки ростков малого бизнеса, инкорпорированного в практику упомянутых технологических гигантов. Они продуктивно размножаются филиалами по всему миру. А для удобства запуска и реализации проектов около крупных предприятий создаются целые «рои» работающих на них малых фирм по схеме так называемых *Local bandit companies*. Это способствует сохранению интеллектуальной собственности новаторов и существенно повышает результативность их творческого труда. Западные университеты и фирмы кропотливо собирают талантливых ученых и перспективные разработки по всему миру. При университетах работают фонды (эндаументы) поддержки науки и образования, существующие за счет безналоговых пожертвований от олигархов и корпораций. Эти фонды, кстати, занимаются селекцией молодых талантов и в других странах, весьма успешно осуществляя внешнюю коммерческую деятельность, наращивая свои капиталы.

Опыт новых технологий бывает не всегда удачным с экономической точки зрения. Но пробовать и начинать – это свойство, как говорят сейчас, «творцов своего счастья». Так, нобелевский лауреат У. Шокли, получивший премию за изобретение транзистора, ознаменовавшего своим появлением микроэлектронную эру, понимал важность коммерциализации своего открытия. Он основал фирму, но обанкротился. Однако начатое им дело выжило. Его коллеге Г. Тилу повезло больше, он нашел способ снижения цены, используя массовое производство транзисторов, и добился успеха.

Подобных примеров можно было бы привести много. И сегодня большинство развитых и развивающихся стран мечтают повторить успешный опыт подобного технологического ренессанса. Поэтому, по мнению многих исследователей, указанный пример информационной революции уже запустил процесс интеллектуальной революции с почти фантастическим использованием информационных технологий, доступной каждому человеку [9-15]. Это явление можно отнести к экономике знаний[5]. А в образовании – это помогло найти новую парадигму смыслов и таких феноменов, как «умные технологии», «умная экономика» и прочее лавинообразное «поумнение» целых областей творческой деятельности и хай-тека. Индикаторами происходящего, как и прежде, становятся новые формы социальной организации труда и вновь создаваемые инструменты глобальных финансов, среди которых некоторые экономисты выделяют область «темной материи экономики и глобальных финансов» [3, 16, 19]. Этот последний феномен, видимо, связан с тем, что рынки ценных и корпоративных бумаг, деривативы, фьючерсы, опционы, форварды, процентные свопы и другие переходят или уже перешли в область глобальных отсроченных долгов, которые, образно говоря, залезают в карман будущих поколений человечества. За это, конечно же, придется платить нашим потомкам. Особенно это касается экономики США с ее раздутым до 26 трлн долл. (на 12 июня 2020 г.) объемом долгов всему миру.

И, если мы считаем, что основной характеристикой современной интеллектуальной экономики является «цивилизация малого параметра» – так называемый глобальный конгломерат кустарей-одиночек, капитализирующих интеллектуальный капитал в будущее, то надо отметить, что этот капитал требует освоения, в том числе, и вне каких-либо государственных границ. А, следовательно, за долги даже самой интеллектуальной части корпораций тех же США (т.е. за долги инноваторов, берущих взаймы у других субъектов глобальной цивилизации) придется расплачиваться не только возможным банкротам, но и всей оставшейся части человечества, если они будут надувать очередные финансовые пузыри таких проектов.

С углублением специализации и разделения современного труда, при делении на передовиков-индивидуалов и прочий отставший индустриальный

[5] Черный В.В., Волков К.А., Есаян Л.Р., Дыбов В.А. Гражданское общество в России. Бизнес и безопасность. – Москва: Агентство безопасности по инвестициям и бизнесу в России, SAIBR, 2006. – 160 с.

мир (в лице традиционных кузниц вроде Китая и стран-тигров), новые индивидуалы, в принципе, могут объединяться в профсоюзы, а профсоюзы – в общины. Но это будут уже общины нового качества. Преимущественным источником создания богатства будет информация и повышенная актуализация процесса реализации новых знаний. Интеллект становится самым востребованным товаром из-за способности приносить сверхприбыль. Поскольку информация в такой системе способна порождать богатство, то преуспевать будут те индивиды, которые выявляют и решают новые проблемы.

ГЛОБАЛИЗАЦИЯ ДЕМОКРАТИИ РАЗВИТИЯ

Процесс описанных нами политико-экономических преобразований сопровождается глобализацией демократических принципов [3, 14, 16]. Это происходит на основе совершенствования культуры освоения информации человеком, увеличения общего объема накопленных им знаний, либерализации информации и финансовых инструментов развивающихся рынков. Естественно, здесь помогает «экономика малого параметра» – тот малый бизнес, который побуждал в людях свободу самовыражения через интеллектуальный ресурс в сумме с предпринимательством. При этом государство под эгидой сильной системы образования и науки, укрепляет суверенитет и безопасность создаваемых систем, стимулирует их выживаемость.

Начиная с интеллектуальных «варягов», процесс миграции по странам ученых и специалистов высоких технологий породил новые проблемы. Политики стараются воспользоваться этим обстоятельством для решения собственных геополитических проблем. А всколыхнувшиеся темные силы капитала пользуются международным разделением труда и глобализацией для использования самой дешевой рабочей силы. С одной стороны, – это воскресило Китай как мировую фабрику ширпотреба, что насытило страны дешевыми товарами. Но в то же время позволило закрывать глаза на несоблюдение норм цивилизованного трудового законодательства. Культуру отношений в этом контексте стали рассматривать как обновленное виртуалом глобальное социальное движение, приведшее к мощным тектоническим сдвигам. Богатые богатеют как в отдельных странах, так и на целых континентах. Разрыв бедности стал усиливаться пугающими темпами, грозящими социальным напряжением. Это неизбежно приводит к перераспределению капитала, его концентрации, и создает предпосылки новых волн социально-экономических кризисов.

Возникшая ситуация требует новых креативных, нестандартных решений. К сожалению, в неразберихе экспертных брожений мы пока не можем обрисовать для России четкую стратегию реструктуризации ее экономики и прописать детали инновационных преобразований. Да и на практике прогнозы сбываются плохо. За такое «неустойчивое равновесие», за свое неумение владеть формами интеллектуальной (да и просто эффективно работающей) собственности, пригодной для технологических прорывов, нам приходится расплачиваться финансово-экономическими потерями.

ЦИФРОВАЯ ЭКОНОМИКА ЗНАНИЙ

Теперь встает вопрос: а ясно ли мы видим (особенно в условиях санкций Запада для России) безоблачные картины своего гармоничного развития? Увы, приходится признать, что мы совсем недалеко ушли от определения классика – В.И. Ленина, который завещал нам учиться, учиться и учиться… . Хотя, если цитировать точно, то эта фраза имеет важнейшее продолжение и рекомендацию: «учиться торговать у капиталистов». Это необходимо помнить, чтобы не повторять случаи, подобные тому, когда Россия в 1993 г. фактически «подарила» 500 тонн оружейного урана для использования в качестве топлива на американских АЭС. А ведь учиться надо не только торговле, но и продуманной долгосрочной конкуренции во всех сферах экономических приоритетов – как высокотехнологических, формирующих уровень цивилизационного развития, так и гуманитарных, определяющих идеологию вектора развития страны.

Это же относится и к не совсем понятной ситуации с решением о предварительном опубликовании материалов диссертационных работ с надеждой на возможное сотрудничество с внешним высокотехнологическим миром. В жесткой борьбе с плагиатом, после всенародного прочтения в открытом доступе и обсуждения содержащихся в этих материалах открытий и научных достижений, теряется коммерческая привлекательность интеллектуальной собственности исследования. Все новое опубликованное, с не защищенной как положено интеллектуальной собственностью, моментально реквизируется опытными конкурентами в мировой гонке за научные открытия и становится старым. Получается, что иногда мы превращаем свою науку и технологии в интеллектуального донора для экономически успешных стран.

Известно, что наука и научные открытия – это движущая сила прогресса. Научные открытия определяют имперское могущество государства на международной арене. Говорят, что «в науке нет ничего практичнее хорошей теории», однако «в науке то ядреное, что в практику внедренное». Поэтому необходимо умело коммерциализировать научный результат для извлечения прибыли, компенсации затрат на научные исследования и создание производства нового продукта. Это же относится и к научно-технологическим достижениям оборонно-промышленного комплекса (ОПК). Пока еще такая коммерциализация является «ахиллесовой пятой» экономики России. Проблема эта очень непростая, но она исключительно важна. Теоретически необходим эволюцион-ный переход от положений добавленной стоимости А. Смита и К. Маркса к добавленной стоимости современных национальных счетов. Практический опыт успешного перехода к такой коммерциализации можно позаимствовать в США и ведущих экономиках Европы. И надо не забывать, что первым в мире обратил внимание на коммерциализацию знаний и торговлю умом россиянин И.Т. Посошков в своем социально-экономическом трактате «Книга о скудости и богатстве» в 1724 г., когда А. Смит еще только родился. Для успешной реализации программы коммерциализации научных достижений и увеличения валовой добавленной стоимости высокотехнологичной и наукоемкой продукции, а также товаров и услуг, в России необходимо совер-

шенствование до современного уровня взаимодействия ОПК, ФСО, РАН, Минэкономразвития, Миннауки, Минпромторга и других ведомств.

Как тут не вспомнить великого М.В. Ломоносова, который еще в 1761 г. призывал: «Размножить миром нашу славу/ И выше как военный звук/ Поставить красоту наук». Вряд ли целесообразно было бы довести науку и технологии до ситуации, аналогичной приснопамятной абсурдной кампании по борьбе с алкоголизмом в СССР 1985–1987 гг. Тогда вырубили уникальные, коллекционные сорта винограда, а сегодня ослабляем поросль поредевших научных школ, отправляя на экспорт талантливых молодых ученых и уникальные плоды их творчества.

Эффективнее было бы, не на словах, а на деле грамотно выстраивать не только кредитно-инвестиционную политику, но и заимствования в экономике и социальной организации труда у Запада, где уже обнаружили, обозначили и применили слагаемые дальнейшего развития и составные части успешных прорывов. В сложившейся ситуации, при обозначенных здесь процессах глобализации демократии, нам, естественно, необходимо не только защищать свои достижения, но и строить стратегию инноваций по законам современного «информационного общества» на основе свободы предпринимательства и умножения главного интеллектуального ресурса в виде человеческого капитала [8]. Подобный подход обеспечит преодоление санкционных издержек России, и ее успешную интеграцию в мировую экономику. Это же касается и регионов страны, которые должны стать продолжением таких инфраструктурных гиперпроектов, как новый космодром «Восточный» и успешные преобразования в оборонно-промышленном комплексе с коммерциализацией инновационных и новаторских решений, а также другие аналогичные драйверы роста, которые пока еще остаются в дефиците для массового применения.

ПРОБЛЕМЫ ЦИФРОВИЗАЦИИ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Председатель правительства РФ М.В. Мишустин [19] считает, что неотрегулированная цифровизация приводит к исчезновению целых секторов экономики вместе с предприятиями и рабочими местами. Она изменяет социальное поведение людей, воздействует на трудовые отношения, на отношение к собственности. Размывается налоговая база, возникают угрозы существованию регуляторов, в том числе общественных и государственных институтов. Происходит исчезновение денежных потоков, поскольку владельцами данных, как товара, становятся большие цифровые компании. А страна, где была создана эта стоимость, теряет прибыль. Государству необходима собственная система цифровой безопасности, инвентаризация всех технологических возможностей и создание центров компетенции в области технологий. В дополнение – электронная демократия и цифровая экономика знаний должны способствовать решению сложных общественных проблем. Разумеется, для развития искусственного интеллекта необходимо развивать и естественный интеллект. Образование в условиях электронной демократии и цифровой экономики знаний должно формировать как глубоко мыслящего профессионала, так и гражданина, осознающего свое место в жизни и процессы, происходящие в государстве и в мире, способного без опоры на пропаганду самостоятельно отличать добро от зла и принимать прогрессивные решения.

Происходящее сегодня из-за пандемии коронавируса общение в формате онлайн способно изменить и парадигму политических процессов. Прежняя форма парламентских и партийных институтов может принять массовый, публичный характер с переходом в неформальный обмен мнениями специалистов и слушателей на видеоконференциях. Особую роль здесь приобретает проблема конфиденциальности данных граждан и неприкосновенность их частной жизни как основы прав человека.

Угрозы национальной безопасности России в основном связаны с цифровизацией национальной экономики по западным рецептам [20]. Поэтому требуется разработка программы проактивного искусственного интеллекта, реализуемого на отечественной технологической платформе [21]. В России есть силы, способные реализовать альтернативную стратегию конструирования будущего для роста общественного блага, в том числе со странами – партнерами по ЕАЭС, которые имеют уникальный опыт живого планирования «затраты-выпуск» с учетом обратной связи и знаний экономической кибернетики для расчета траектории движения к поставленной цели, т. е. конструирование будущего в желаемом направлении.

ЗАКЛЮЧЕНИЕ

Систему проактивного искусственного интеллекта [21] целесообразно выстроить на основе динамической модели межотраслевого межсекторного баланса, организующего информационные потоки от экономических объектов для координации их деятельности в направлении реализации стратегических целей и ускорения темпа движения за счет внедрения новых технологических способов производства. Необходимо создавать динамические экономические модели, в которых должны отражаться механизмы действия и реализации объективных экономических законов развития [22].

Чем больше «новой» собственности, тем больше гражданского общества – этот вывод следует из обзора технологических революций, сопровождающих виртуальное и интеллектуальное развитие, продемонстрированное развитыми странами [14-17]. А в России, особенно в ее управленческом звене и среди элиты, все-таки необходимо развивать осознание тех фактов, что только через эффективные формы собственности и цивилизованный рынок может родиться успешный средний класс, а заодно и социально активный, работающий инноватор, генератор собственных уникальных идей. Поддержка и защита всех форм собственности, включая интеллектуальную, требует развития, а может быть даже и качественного скачка в правосознании всего общества, которое бы могло отречься от примитивных, в том числе олигархических способов существования. В этом смысле, например, интересно выглядит предложение Газпрома продавать акции компании всем желающим. Но для реализации подобной идеи нужен растущий платежеспособный спрос населения.

В обществе должен нарастать и развиваться здоровый личный интерес к освоению наукоемких производств и всего хай-тека. Понадобится настоящий моз-

говой штурм со стороны прогрессивной части россиян. Миру нужна сильная в научно-технологическом отношении Россия с развитой диверсифицированной экономикой, отвечающей нуждам населения. Так сложилось исторически – без России мир не полон.

СПИСОК ЛИТЕРАТУРЫ

1. Бажанов Е.П., Бажанова Н.Е. Куда идет человечество? О тенденциях международных отношений в XXI веке. – М.: Восток – Запад, 2009.
2. Бажанов Е.П. Россия и Запад // Международная жизнь. – 2013. – № 12. – С. 11-36.
3. Бажанов Е.П. Глобализация как объективный процесс // Эхо планеты. – 2010. – № 32(27 авг.– 2 сент.) – С. 21.
4. Добрецов Н.Л., Золотов Ю.А., Иванов В.Т., Леонтьев Л.И., Макаров А.А., Мясоедов Б.Ф., Наточин Ю.В., Островский М.А., Розанов А.Ю., Ушачев И.Г., Черноусько Ф.Л. Совет старейшин Российской академии наук: Реформа — удар по российской академической науке. Меры по повышению роли РАН в научно-технологическом развитии России // Представительная власть – XXI век. – 2020. – № 1-2. – С. 12-20.
5. Смирнов А.А. Обеспечение информационной безопасности в условиях виртуализации общества: опыт Европейского Союза. – М.: ЮНИТИ-ДАНА, 2011.
6. Cherny V.V., Kapkov A.Yu. Russia and the USA: virtual games of superpowers (The “end of history” is canceled, and “the clash of civilizations” lead to new models of democracy, international cooperation, and international security) // European Security. Frank Cass. USA. – 2000. – Vol. 9, № 3. – P. 123-133.
7. Любимов А.П. Административная и правовая основа электронной демократии // Представительная власть – XXI век. – 2012. – № 4. – С. 2-4.
8. Любимов А.П. Перспективы создания российских инновационных кластеров // Представительная власть – XXI век. – 2013. – № 5-6. – С. 14-19.
9. Любимов А.П. Об общественном (публичном) контроле за компьютерным подсчетом голосов во время выборов // Законодательство. – 1998. – № 11. – С. 69-76.
10. Бабкин В.В., Промоненков В.К., Овчаренко М.М., Любимов А.П. Инновационная концепция средств защиты растений в Российской Федерации // Химическая промышленность сегодня. – 2017. – № 8. – С. 50-54.
11. Любимов А.П. Основные подходы к определению понятия «искусственный интеллект» // Научно-техническая информация. Сер. 2. – 2020. – № 9. – С. 1-6.
12. Любимов А.П. Достоинства и недочеты двух важных законопроектов: мнения экспертов. Круглый стол в Государственной Думе // Журнал российского права. – 2000. – № 4. – С. 26-27.
13. Любимов А.П., Пономарева Д.В., Барабашев А.Г. К вопросу о понятии искусственного интеллекта в российском праве // Актуальные вопросы экономики, управления и права: сборник научных трудов (ежегодник). – 2019. – № 2-3.– С. 16-34.
14. Ложковой П.Н. Правовая природа деятельности государств по дистанционному зондированию Земли из Космоса // Актуальные вопросы экономики, управления и права: сборник научных трудов (ежегодник). – 2018. – № 4. – С. 4-19.
15. Pronchev G.B., Lyubimov A.P., Proncheva N.G., Tretiakova, I.V. Social and economic causes of labor migration in contemporary Russia // Espacios. – 2019. – T. 40, № 32. – С. 13.
16. Капков А.Ю., Черный В.В. Эволюция глобализма: от коллективного бессознательного – к коллективному сознательному. – М.: Изд-во С.П. Шукшиной, 2014. – 202 с.
17. Pronchev G.B., Mikhailov A.P., Lyubimov A.P., Solovyev A.A. Particularities of the Internet-based virtual social environments within the context of information warfare // EurAsian Journal of BioSciences. – 2020. – Vol. 14. – P. 3731 – 3739.
18. Щитов А.Н. Проект «Открытая наука России» // Актуальные вопросы экономики, управления и права: сборник научных трудов (ежегодник). – 2020. – № 2-3. – С. 77-86.
19. Мишустин М.В. Построение устойчивого региона на основе данных и искусственного интеллекта // Международный форум «Цифровое будущее глобальной экономики» (31.01.2020. Алма-Ата, Казахстан). – URL: http://government.ru/news/38885/ (дата обращения 10.11.2020 г.)
20. Ведута Е.Н. Цифровая экономика приведет к экономической киберсистеме // Международная жизнь. – № 10. – 2017. – С. 87-102.
21. Ведута Е.Н., Любимов А.П., Джакубова Т.Н., Ряскова Е.С. Концепция национальной программы создания проактивного искусственного интеллекта // Представительная власть – XXI век. – 2019. – № 4. – С. 22-29.
22. Ведута Е.Н. Стратегия и экономическая политика государства. – М.: Академический проспект, 2004. – 456 с.

Материал поступил в редакцию 19.11.20.

Сведения об авторах

ЛЮБИМОВ Алексей Павлович – доктор юридических наук, профессор, заместитель Главного ученого секретаря Президиума РАН, Москва
e-mail: aplyubimov@presidium.ras.ru

ЧЕРНЫЙ Владимир Викторович – доктор физико-математических наук, вице-президент Фонда защиты конституционных прав коренных малочисленных народов России, член коллегии Российского агентства развития информационного общества, член Союза журналистов России, Москва.
e-mail: chernyv@list.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК [004.89:005.521] : 001.89

П.А. Калачихин

Обоснование показателей для управления научными достижениями*

Рассматривается проблема обоснования оптимального состава показателей, предназначенных для оценки достигнутых и прогнозирования новых научных достижений. Систематизируются типы показателей, которые обычно используются в управлении научными достижениями. Предлагается дифференцированный подход к выбору таких показателей в зависимости от разделов знания, к которым они относятся. Помимо наукометрических параметров разделов знания, перечисляются факторы, оказывающие влияние на формирование наборов показателей. Представлена разработка количественной модели соотношения типов показателей в составе их наборов на основе мер множеств и бинарных отношений порядка над числами. В рамках этой модели дается объяснение превалированию экспертных показателей. Решение о составе наборов показателей принимается на основании эвристических правил. Дается пример поиска оптимального соотношения типов показателей для прогнозирования достижений естественных наук и оценки достигнутых результатов гуманитарных наук.

Ключевые слова: *выбор показателей, гуманитарные науки, естественные науки, набор показателей, научное достижение, оценка результативности, экспертное прогнозирование*

DOI: 10.36535/0548-0027-2021-03-2

ВВЕДЕНИЕ

Убеждение, что наука – это особая, в некотором роде, «священная» сфера, эффективно управлять которой в состоянии только «жрецы», т.е. сами ученые, было распространено ранее. Казалось бы, ученые должны оценивать результаты своей работы так, как им больше всего нравится. Соответственно, критерии оценки и прогнозирования научной результативности требуют творческого подхода и поэтому должны разрабатываться самими же учеными.

Однако со временем в руководство отечественной науки стали приходить управленцы – менеджеры. Согласно теории менеджмента, прогнозирование научных достижений относится к планированию научной деятельности, а оценку научных результатов следует отнести к контролю за научной деятельностью. В свою очередь, планирование и контроль разбиваются на множество более мелких задач, каждая из которых требует отдельного внимания.

Современная школа научного менеджмента, ориентированная на учет национальных интересов, более системно относится к выбору критериев управления, фокусируясь на менее творческих, но более надежных методиках. Глобальная установка на унификацию наборов показателей для всех отечественных научных организаций заставляет еще выше поднимать планку требований к методологии отбора показателей.

В настоящем исследовании мы ставим перед собой задачу поиска оптимального соотношения показателей разных типов в составе наборов показателей, используемых для оценки результатов уже полученных и прогнозирования новых научных достижений. Под *научными достижениями* (*scientific achievements*) будем понимать результаты фундаментальных исследований и прикладных разработок научно-исследовательских, научно-инновационных и научно-образовательных организаций. При этом условимся полагать, что прогнозирование осуществляется в среднесрочном периоде и затрагивает как отдельные результаты интеллектуальной деятельности, так и научные направления и категории знания в целом.

* Работа выполнена в рамках исследования по теме 0003-2019-0001 Госзадания ВИНИТИ РАН и при поддержке Российского фонда фундаментальных исследований (проект РФФИ № 20-07-00014).

ТИПОЛОГИЯ ПОКАЗАТЕЛЕЙ, ПРИМЕНЯЕМЫХ В ЗАДАЧАХ ПО УПРАВЛЕНИЮ НАУЧНЫМИ ДОСТИЖЕНИЯМИ

Использование наукометрических методов представляется перспективным для решения, например, таких задач:

- анализ структуры и уровня отечественной и мировой науки;
- определение тенденций и процессов, происходящих в мировой и региональной науке;
- выявление (на ранней стадии) наиболее актуальных или, напротив, теряющих свою актуальность научных направлений;
- отслеживание генезиса конкретных научных идей (или направлений) и истории их развития;
- определение продуктивности научных организаций и работы отдельных исследователей (научных групп) в конкретной научной области и эффективности материальных и иных затрат в этой области;
- изучение трендов развития инновационной деятельности в рамках отдельных научных организаций, направлений (или отделений РАН);
- исследование структуры научного сообщества и изучение науки как социального организма [1, с. 123].

Некоторые из этих задач условно можно отнести к оценке достигнутых научных результатов, а другие – к прогнозированию новых достижений. Так, определение продуктивности научной работы относится к подвиду задач, связанных с оценкой результативности; выявление наиболее актуальных направлений – к задачам прогнозирования.

Многие типы показателей, используемые в таких задачах, устроены иерархически, поэтому для того, чтобы классифицировать и систематизировать основную их часть, воспользуемся представленной на рис. 1 таксономией.

Существует два блока инфометрических показателей: традиционные (наукометрические и библиометрические) и сетевые в составе вебометрик и альтметрик. При этом *инфометрические* показатели, которые также называют *инфометриками*, могут извлекаться из результатов параметризованных запросов к наукометрическим базам данных. Таким образом, инфометрические показатели необходимо расширить показателями, дополненными текстовыми данными в рамках семантических технологий [2].

В то время, как инфометрические показатели по праву относят к формальным, экспертные показатели (*expert indicators*) подсчитываются на основании субъективного экспертного мнения. Активное применение экспертных показателей в управлении наукой имеет под собой веские основания.

Не нужно полагаться только на одни формальные показатели, так как не следует исключать возможность существования в науке таких интересных явлений, как «спящая красавица» (*sleeping beauty*), т. е. статья, опубликованная много лет назад и получившая «взрыв» цитирований в настоящее время [3], и «черный лебедь» (*black swan*), т. е. событие, которое является неожиданными и влечет за собой значительные последствия, хотя имеет рациональное объяснение [4].

Количественная оценка должна дополнять качественную, экспертную оценку. Количественные измерения могут уравновесить возможное предубеждение перед экспертным рецензированием (*peer review*) и упростить обсуждение. Они, как правило, усиливают экспертное рецензирование, поскольку трудно судить коллег, не владея спектром необходимых сведений. Тем не менее, специалисты, оценивающие научную деятельность, не должны следовать соблазну переложить принятие решений на числа. Индикаторы – не замена информированному суждению. Каждый эксперт сохраняет ответственность за собственную оценку [5].

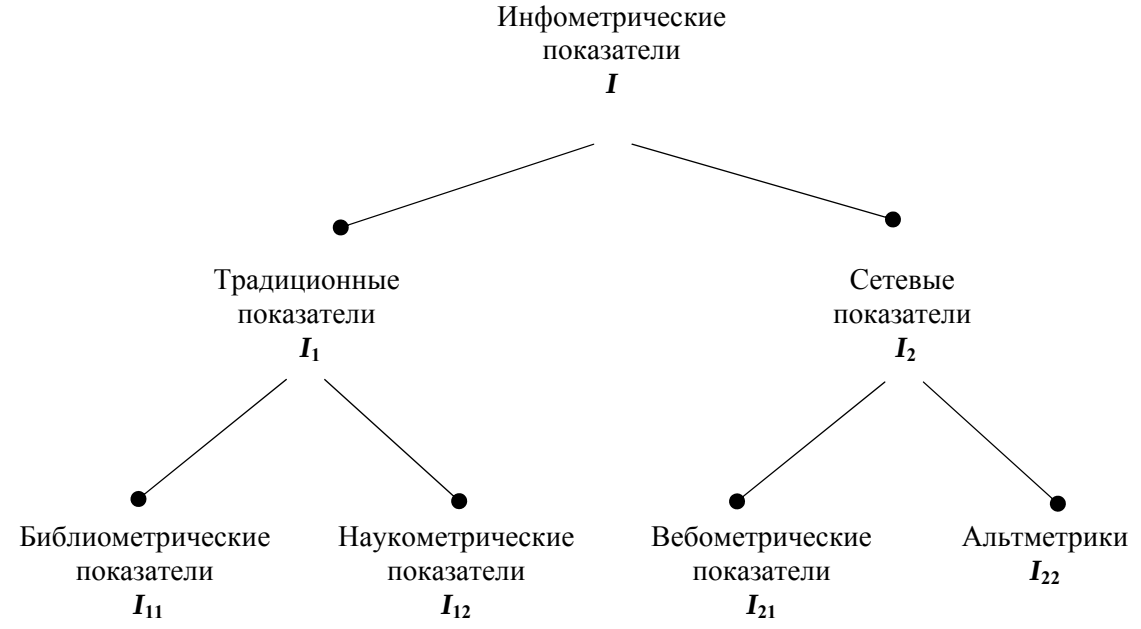

Рис. 1. Таксономия инфометрических показателей, используемых в управлении научными достижениями

В последнее десятилетие значительная часть «научных статей» и «диссертаций» отечественных авторов, особенно в социально-гуманитарных областях, оказывается имитацией науки [6]. Недобросовестность в проведении исследований может приобретать форму научного мошенничества, которое включает фабрикацию, фальсификацию, плагиат и незаконное присвоение чужих результатов. Наряду с этим существуют менее грубые нарушения, которые обозначаются как спорные исследовательские практики [7, с. 30].

В расследовании нарушений научной этики экспертиза играет незаменимую роль, хотя техническая сторона подобной экспертизы, как правило, предполагает лексическую проверку текстов на оригинальность, библиометрическое обнаружение недобросовестного авторства либо применение других формальных методов. Помимо этого, пока что только на основании экспертизы можно отличить так называемую «*девиантную науку*» от *протонауки*, так как подобная демаркация требует уже не формального, а качественного анализа и индивидуального подхода.

Существует принципиальная разница в экспертной оценке и экспертном прогнозировании научной деятельности. Если в первом случае можно говорить о каких-то конкретных показателях, задаваемых теми или иными шкалами, то второй случай выглядит более туманным, и обычные экспертные показатели здесь не годятся. Для экспертного прогнозирования нужно использовать *экспертные технологии*, так как здесь важна именно методология оценки, а не используемые показатели. Основными элементами экспертных технологий являются способы формирования экспертных групп, критерии отбора специалистов в экспертные группы, способы опроса экспертов, процедуры и методы организации совместной деятельности экспертов [8]. Помимо этого, экспертное прогнозирование научной деятельности может включать оценку рисков, которые однозначно содержат элемент прогноза.

В случаях, когда неизвестно, как наилучшим образом подсчитывать заданный показатель и не разработана адекватная методика его оценки, в качестве альтернативы обычно применяются экспертные оценки. Однако не всегда очевидно, что это лучший выход. Возможно, лучше вначале потратиться на методику как в пословице – скупой платит дважды (*stingy always pays twice*).

ПОДБОР ПОКАЗАТЕЛЕЙ ДЛЯ ОЦЕНКИ И ПРОГНОЗИРОВАНИЯ НАУЧНЫХ ДОСТИЖЕНИЙ

В целом, можно согласиться с точкой зрения, которой придерживаются многие отечественные ученые, что критерии оценки результатов научной деятельности должны учитывать специфику разделов знания, в которых они были получены, т. е. для оценки достижений разных наук методологически неверно использовать одни и те же показатели. Следует отметить, что развиваемый нами подход во многом заимствован из Методики расчета качественного показателя государственного задания «Комплексный балл публикационной результативности» для научных организаций, подведомственных Министерству науки и высшего образования Российской Федерации [9]. Первоначальная редакция документа вызвала критику, связанную с тем, что академическому сообществу представилось неправильным оценивать деятельность организаций, занимающихся общественными и гуманитарными науками, по тем же критериям, что и деятельность остальных организаций. Таким образом, дифференцированный подход появился в поздних версиях документа.

Хрестоматийным является пример индекса Хирша и прочих «хиршеподобных» показателей, которые могут быть «накрученными» у авторов, специализирующихся в тех областях, где публикации имеют большое количество соавторов, что характерно для естественных и особенно физических наук. В частности, публикации по физике высоких энергий (*high energy physics*) очень часто имеют большое количество соавторов, поскольку над экспериментальным оборудованием и обработкой данных могут работать сотни специалистов и исследователей.

В свою очередь, представители социальных и гуманитарных наук в среднем имеют меньше статей и заметно реже ссылаются на других авторов, чем это происходит в естественнонаучных дисциплинах и, соответственно, обладают более низкими индексами цитируемости. Естественные науки, с одной стороны, и социальные и гуманитарные науки – с другой, различаются характером публикаций (монографии против журналов), языком, целевой аудиторией и задачами, привычками и практикой цитирования, числом ежегодных публикаций, количеством исследователей в группе и видом авторства (индивидуальным или коллективным) [10].

Помимо этого, нужно обращать внимание на методики расчета показателей и параметры, используемые в этих методиках. Например, когда оцениваются статейные публикации, то импакт-факторы журналов в гуманитарных или социальных науках за 2 года менее полезны, чем за 5 лет, поскольку в целом публикации по социо-гуманитарным наукам цитируются реже, чем по естественным.

Можно полагать, что наиболее современные науки, к которым следует в первую очередь отнести так называемые «вычислительные науки» (*computer science*) и их приложения, лучше поддаются выражению через сетевые показатели. Это может быть связано с новыми технологическими трендами в научных коммуникациях, библиотечной и издательской политике. Как правило, передовые результаты исследований публикуются в тех изданиях, которые применяют наиболее современные компьютерные технологии.

В силу того, что различные разделы знания имеют свою специфику, можно выделить особый тип показателей, которые далее будем называть *специальными* (*special*). Специальный показатель может быть актуален только для одного единственного или ограниченного числа разделов знания. Например, только в сельскохозяйственных науках результаты исследований могут быть подсчитаны через селекционные показатели количества новых сортов, гибридов, пород. Сюда же будем относить показатели, которые не относятся ни к инфометрическим показателям, ни к экспертным оценкам. В частности, для технических наук – это оценки различных видов потенциала и перспектив-

ности разработок, а также показатели, учитывающие коммерциализацию результатов исследований.

Специальные показатели востребованы прежде всего в прикладных науках. В то время как результаты фундаментальных исследований оцениваются наукометрическими методами при помощи таких показателей, как количество цитирований и импакт-факторы журналов, в которых они были опубликованы, результаты прикладных исследований могут быть оценены экономическими методами посредством периода окупаемости, чистого приведенного дохода, индекса прибыльности и других финансовых показателей [11].

Следует отметить, что специфика раздела знания не всегда выражается через специальные показатели, в некоторых случаях достаточно имеющегося стандартного набора. Так, оценка результативности в гуманитарных науках может быть выражена через показатели публикационной активности, дополненные количеством изданных словарей, собраний сочинений, энциклопедий и архивно-документальных публикаций. Для прочих наук такие форматы публикаций являются редким, но все же не исключаемым явлением.

Итак, в нашей информационной модели формальные показатели представлены стандартными инфометрическими показателями, а также специальными показателями, которые могут иметь разную форму.

Экспертная же оценка может осуществляться по вербальным, балльным, интервальным и количественным шкалам [8]. Логично полагать, что раздел знания мало влияет на выбор шкалы для экспертного оценивания. Однако при выборе шкал следует учитывать прочие условия из постановки задачи по управлению научными достижениями. Так, для качественного прогнозирования подходят вербальные (качественные) шкалы, а для количественного – годятся балльные, интервальные и прочие количественные шкалы.

Тем не менее, раздел знания сильно диктует выбор экспертной технологии. Так, некоторые авторы придерживаются точки зрения, что если для естественных наук допустимо формировать экспертные комиссии, помимо всего прочего опираясь на библиометрические показатели кандидатов, то в гуманитарных науках «существуют совершенно иные публикационные традиции, и реализация рекомендательного механизма выбора научных экспертов (если она вообще возможна) должна быть устроена как-то иначе» [12, с. 335].

Можно предположить, что выбор показателей прогнозирования достижений науки так же, как и оценка результативности научной деятельности дифференцируется по разделам знания. В целом прогнозировать сложнее, чем оценивать, так как при прогнозных оценках приходится сталкиваться с большей неопределенностью и, как следствие, – прилагать дополнительные усилия, чтобы добиться точности прогноза. Если утверждать, что для построения количественных прогнозов достаточно формальных показателей, то для качественных прогнозов экспертные оценки жизненно необходимы. Таким образом, состав показателей при оценке результатов и при прогнозировании достижений должен выглядеть по-разному. Однако, рассматривая параметры, влияющие на выбор этих показателей, следовало бы учесть еще ряд факторов.

Известно, что никакая экспертиза невозможна без компетентных экспертов. Помимо этого, экспертиза больших объемов данных требует значительных затрат. В тех ситуациях, когда нет достаточного количества грамотных экспертов, либо нет ресурсов на оплату их труда, вместо того, чтобы снижать качество экспертизы, можно от нее вовсе отказаться, сделав упор на формальные показатели, которые всегда извлекаются из исходных данных.

В случае, когда нет доступа к базам данных, например, ввиду языковых ограничений или платной подписки, приходится возвращаться к экспертным оценкам. То же можно сказать в отношении платформ-провайдеров альтметрик, которые требуют наличия у пользователей некой «сетевой культуры», свойственной далеко не всем авторам. Если нет возможности собирать альтметрики, то следует обращаться к наукометрическим и библиометрическим показателям.

При выборе показателей важную роль играет доверие представителей того или иного научного сообщества к тому или иному их виду:

- к экспертным оценкам (к институту экспертизы, как правило, в конкретном государстве);
- к альтметрикам, цифровым онлайн-технологиям, а также сопротивление или недовольство технологическими инновациям (что может стать препятствием к переходу на сетевые показатели).

Таким образом, параметры различных разделов знания могут содержать информацию о том, в какой степени велика публикационная активность в данном разделе науки в целом, насколько активно цитируются публикации этого раздела и даже как велико воздействие публикаций по данному разделу науки на те или иные сферы деятельности. Тем не менее, выбор показателей не всегда зависит только от наукометрических и родственных им параметров рассматриваемых разделов знания, так как решающую роль могут сыграть другие неожиданные факторы.

МОДЕЛИРОВАНИЕ СООТНОШЕНИЙ ПОКАЗАТЕЛЕЙ В ЗАДАЧАХ ПО УПРАВЛЕНИЮ НАУЧНЫМИ ДОСТИЖЕНИЯМИ

Рассмотрим решение задачи поиска оптимального весового соотношения формальных показателей и экспертных оценок, инфометрических и специальных показателей, типов инфометрических показателей (традиционные – науко- и библио-, сетевые – альт и вебо- метрики).

Прежде всего, следует признать высокую важность всех показателей, которые имеют целевое значение, и тех показателей, которые подлежат оптимизации. В свою очередь, низкой важностью обладают те показатели, которые не актуальны для данного раздела знания, легко накручиваются или порождают аномалии. Остальные показатели, составляющие большинство в общей массе, имеют средний или соседний с ним уровень важности.

Таким образом, чтобы оценить важность того или иного показателя в контексте наборов показателей, формируемых для выполнения той или иной задачи, можно воспользоваться шкалой вида:

{″Очень низкая″; ″Низкая″; ″Средняя″;
″Высокая″; ″Очень высокая″}, (1)

где очень низкой важности соответствует 0 баллов; низкой – 25; средней – 50; высокой – 75; очень высокой – 100 баллов.

В идентифицированной и упорядоченной балльной шкале (1) интервалы между уровнями равны, при этом шкала содержит абсолютный 0 и медианное значение. Тем самым шкала оценки важности показателя имеет достаточно «продвинутый» (*advanced*) вид.

Следует отметить, что для качественной экспертизы требуется ограничить состав экспертной группы, поскольку при излишнем количестве экспертов точность прогноза снижается [13]. Обычно единоличный эксперт – это сомнительная инициатива, но, чтобы определить важность показателей, всего единственный эксперт не является чрезмерной крайностью, поскольку процесс организации экспертизы существенно упрощается. В этой связи вопросы получения агрегированных экспертных оценок далее рассматриваться не будут.

Существуют и другие способы определения важности показателей. Например, путем сопоставления существующего и планируемого уровня значения показателя или на основании ресурсных критериев [14]. Однако мы остановились на экспертизе, поскольку при выполнении однотипной задачи плановые и фактические значения показателей, а также имеющиеся ресурсы от случая к случаю могут различаться, и каждый раз набор показателей нужно было бы пересматривать, в то время как мы желаем добиться некоторой универсальности.

Важности показателей нормируются при помощи преобразования, позволяющего получить относительную важность для отдельного показателя, входящего в состав некоторого набора показателей:

$$a_i^* = \frac{a_i}{\sum a_i}\ , \qquad (2)$$

где a_i^* – нормированная по U важность i-го научного показателя u_i, $u_i \in U$; a_i – важность i-го научного показателя u_i; U – множество показателей из набора. При этом выполняется нормирующее условие $\sum a_i = 1$.

В свою очередь, важность показателей α-го типа рассчитывается следующим образом:

$$A_\alpha^* = \sum_j a_{\alpha,j}^* \qquad (3)$$

где A_α^* – суммарная нормированная по множеству показателей из набора U важность показателей α-го типа; $a_{\alpha,j}^*$ – нормированная по U важность j-го показателя α-го типа $u_{\alpha,j}$, $u_{\alpha,j} \in U_\alpha$; U_α – множество показателей α-го типа из набора показателей U. При этом выполняется нормирующее условие $\sum A_\alpha^* = 1$.

Важность отдельного показателя и суммарную важность набора показателей заданного типа также можно определить через меру w множества X – неотрицательную функцию $w(X) \geq 0$, определенную на семействе множеств $\mathcal{F}$ и удовлетворяющую условию аддитивности:

$$w(X \cup Y) = w(X) + w(Y), \qquad (4)$$

где $X, Y \in \mathcal{F}$; $X \cap Y = \varnothing$.

При этом мера пустого множества равна нулю:

$$w(\{\varnothing\}) = 0 \qquad (5)$$

В таком случае важность одного показателя является мерой множества показателей, состоящего из единственного элемента:

$$a_i = w(\{u_i\}) \qquad (6)$$

где: a_i – важность i-го научного показателя; u_i – i-й научный показатель в наборе U; w – мера, заданная на множествах (наборах) научных показателей.

В свою очередь, суммарная важность показателей типа α является мерой множества показателей заданного типа:

$$A_\alpha = w(U_\alpha), \qquad (7)$$

где A_α – суммарная важность множества показателей α-го типа; U_α – множество показателей α-го типа; w – мера, заданная на множествах показателей.

Распределение по табл. 1 мы получим, если поставить все показатели, что называется, в равные условия (хотя на практике, как правило, все складывается по-другому).

Коэффициенты относительной важности типов показателей должны подсчитываться с одинаковой точностью, которая обеспечивается достаточным количеством знаков после запятой. Поскольку у наукометрических показателей относительная важность составляет *0,0625*, а у экспертных показателей – *0,5*, то главный (преобладающий) тип – это экспертные показатели.

На рис. 2 крупные типы показателей распределены по трем блокам.

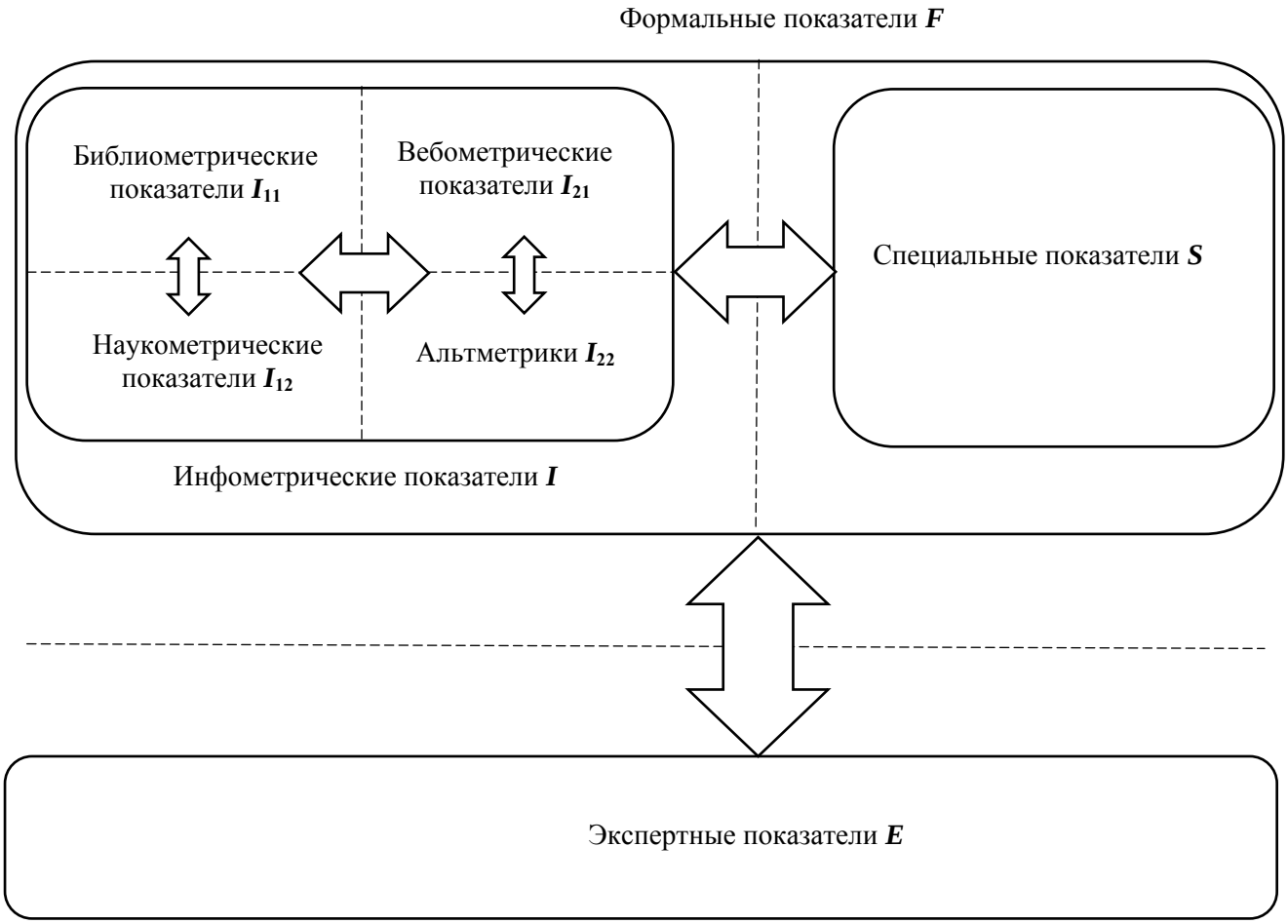

Рис. 2. Схема соотношений между разными типами показателей.
(Пунктирные линии показывают границы между типами показателей.
Двойные стрелки соответствуют соотношениям внутри 5-ти пар типов показателей).

Таблица 1

Распределение коэффициентов относительной важности типов показателей по умолчанию

Коэффициент относительной важности типа показателей	Тип показателя	Значение показателя по умолчанию
A_F^*	Формальный	0,5
A_E^*	Экспертный	0,5
A_I^*	Инфометрический	0,25
A_S^*	Специальный	0,25
$A_{I_1}^*$	Традиционный	0,125
$A_{I_2}^*$	Сетевой	0,125
$A_{I_{11}}^*$	Библиометрическй	0,0625
$A_{I_{12}}^*$	Наукометрический	0,0625
$A_{I_{21}}^*$	Вебометрический	0,0625
$A_{I_{22}}^*$	Альтметрики	0,0625

Наш набор показателей должен быть максимально компактным:

$$N \to \min \quad (8)$$

где N – суммарное количество используемых научных показателей в наборе.

Это значит, что из двух наборов U_α и U_β с одинаковыми важностями $w(U_\alpha) = w(U_\beta)$ предпочтительнее тот набор, в котором меньше количество показателей *N*. Из этого следует, что показатели с очень низкой важностью, которые согласно шкале (1) оцениваются величиной 0 баллов, нужно исключать.

Следует учитывать, что формальные показатели могут быть как элементарными, т. е. простыми, так и композитными, т. е. сложными. Каждый компонент сложного формального показателя должен принадлежать одному из типов показателей (инфометрическому или специальному), либо рекурсивно быть многокомпонентным. В таком случае наша задача сводится к подсчету весовых долей компонентов разных типов в составе сложных показателей. Если объединение компонентов осуществляется линейной сверткой, то достаточно подобрать оптимальные весовые коэффициенты.

Как мы видим, набор простых показателей подменяется набором из компонентов сложного показателя. Использование таких многокомпонентных показателей допускается, хотя и не рекомендуется, наряду с непрозрачными коэффициентами и сомнительными рейтингами. Тем не менее, для композитных показателей действует та же самая модель соотношения типов, что и для обычных показателей. Аналогично и относительно гибридных показателей, сочетающих формальные показатели (простые или сложные) с экспертными оценками.

Различия между композитными и гибридными показателями заключаются в способах сочетания компонентов, которые будем называть *композитными технологиями*, когда сочетаются только формальные показатели, *и гибридными*, если в состав сочетаемых компонентов входят экспертные показатели. Как правило, способ сочетания компонентов в композитных показателях представлен целевой функцией, например, как в [15]. Гибридные технологии имеют более разнообразный вид, в частности, это могут быть методы ранжирования [16]. В качестве универсального метода агрегирования выступают нечеткие меры и интегралы [17]. На наш взгляд, композитные и гибридные показатели являются недостаточно проработанной темой исследования.

ЭВРИСТИЧЕСКИЕ ПРАВИЛА ПРИНЯТИЯ РЕШЕНИЙ О СОСТАВЕ ПОКАЗАТЕЛЕЙ ДЛЯ УПРАВЛЕНИЯ НАУЧНЫМИ ДОСТИЖЕНИЯМИ

Непосредственно для целей нашего исследования необходимо формализовать экспертное знание, чтобы стало возможным сопоставлять по относительной важности типы показателей в составе наборов. Согласно применяемому в библиометрии эвристическому подходу (*heuristic approach*), решения по сопоставлению объектов могут приниматься на основании так называемых *эвристик* (*heuristics*), которые в теории принятия решений (*decision theory*) принято называть *правилами принятия решений* (*decision rules*) [18]. Получение точных рекомендаций по целевым значениям относительных важностей типов показателей проблематично, поэтому рекомендации пока что могут принимать лишь вид бинарных отношений порядка над числами. При этом, чем большей информацией о начальных наукометрических параметрах разделов знания мы владеем, тем детальнее становится возможность описывать соотношения, т. е. получать более «сильные» (строгие) бинарные отношения порядка относительно большего количества типов показателей. Вообще, соотношение внутри части знания должно наследовать ограничения из вышестоящего раздела знания.

Нужно учитывать, что на данный момент мнение научного сообщества по наукометрическому вопросу далеко не однозначно. Более того, имеются полярные точки зрения. При этом наиболее активные позиции занимают ярые противники использования библиометрии и наукометрии, так же проявляют себя и скептики в отношении ценности альтметрик. Вероятно, что решение этого вопроса будет зависеть от того, чьи позиции на поле научных дебатов сильнее, или какие показатели сейчас в тренде у научного сообщества.

Правила, связанные с выбором показателей для построения среднесрочных количественных прогнозов в естественных науках, могут выглядеть следующим образом:

1. **IF** Задача = "*Прогнозирование*" **AND** Вид прогноза = "*Количественный*" **THEN** $A_F^* > A_E^*$;

2. **IF** Задача = "*Прогнозирование*" **AND** Область знания = "*Естественные науки*" **THEN** $A_I^* < A_S^*$.

Собственно, эти правила предназначаются, чтобы смещать соотношения, взятые по умолчанию из табл. 1, исходя из практических соображений. Однако, применяя их мы получим систему неравенств, которую можно использовать в дальнейшем. Например, если при прогнозировании ограничиться детализацией до области знания, то система неравенств для естественных наук могла бы выглядеть так:

$$\begin{cases} A_F^* > A_E^* \\ A_I^* < A_S^* \\ A_{I_1}^* \; R \; A_{I_2}^* \\ A_{I_{11}}^* \; R \; A_{I_{12}}^* \\ A_{I_{21}}^* \; R \; A_{I_{22}}^* \end{cases} \quad (9)$$

где: A_F^* – коэффициент относительной важности формальных показателей; A_E^* – коэффициент относительной важности экспертных показателей; A_I^* – коэффициент относительной важности инфометрических показателей; A_S^* – коэффициент относительной

важности специальных показателей; $A^*_{I_1}$ – коэффициент относительной важности традиционных показателей; $A^*_{I_2}$ – коэффициент относительной важности сетевых показателей; $A^*_{I_{11}}$ – коэффициент относительной важности библиометрических показателей; $A^*_{I_{12}}$ – коэффициент относительной важности наукометрических показателей; $A^*_{I_{21}}$ – коэффициент относительной важности вебометрических показателей; $A^*_{I_{22}}$ – коэффициент относительной важности альтметрик; R – бинарное отношение порядка над числами (много больше $\gg$, больше $>$, больше или равно $\geq$, приблизительно равно $\approx$, меньше или равно $\leq$, меньше $<$, много меньше $\ll$).

Рассмотрим другой пример эвристического правила выбора показателей для оценивания результативности исследований в гуманитарных науках:

3. **IF** Задача = "*Оценка*" **AND** Область знания = "*Гуманитарные науки*" **THEN** $A^*_F < A^*_E$ **AND** $A^*_I > A^*_S$.

Для оценки результатов исследований в гуманитарных науках система неравенств могла бы выглядеть так:

$$\begin{cases} A^*_F < A^*_E \\ A^*_I > A^*_S \\ A^*_{I_1} \; R \; A^*_{I_2} \\ A^*_{I_{11}} \; R \; A^*_{I_{12}} \\ A^*_{I_{21}} \; R \; A^*_{I_{22}} \end{cases} \qquad (10)$$

где обозначения совпадают с формулой (9).

Таким образом, в приведенных примерах рекомендации для оценки результативности в гуманитарных науках противоположены рекомендациям для количественного прогнозирования достижений естественных наук, если детализировать соотношение показателей с точностью до области знания. Для того, чтобы получить более детальные рекомендации, нужно привлечь информацию о том, к какому направлению или категории относятся оцениваемые достигнутые научные результаты или прогнозируемые достижения научной деятельности.

ЗАКЛЮЧЕНИЕ

В связи с тем, что среди факторов, влияющих на формирование наборов показателей результатов научных исследований, можно выделить формальные параметры разделов знания, а также начальные условия постановки задач, дифференцированный подход к оценке достигнутых результатов научной деятельности и прогнозированию новых научных достижений представляется нам уместным. Однако было показано, что на выбор таких показателей могут оказывать заметное влияние многие внешние факторы: ограничение по ресурсам, доступ к технологиям, доверие к технологиям и институтам, наличие лингвистического или культурного барьеров и прочее.

Наукометрические показатели имеют более глубокую нишу в таксономии показателей, поэтому, по умолчанию, обладают меньшей относительной важностью в общем соотношении. Следовательно, использование наукометрических показателей в той же мере, что и экспертных оценок, равнозначно переносу центра тяжести на наукометрические показатели.

Передача оценки важности показателей экспертам на деле может проявить себя отрицательным образом, так как повлечет смещение оценки в сторону экспертных показателей. Помимо этого, подсчет коэффициентов относительной важности показателей может потребовать большей точности вычислений.

Эвристические правила, формализующие экспертное знание о том, какие типы показателей рекомендуется использовать, играют ключевую роль в определении типового состава показателей для упомянутых примеров прогнозирования новых научных достижений в естественных науках и оценки достигнутых результатов научной деятельности в гуманитарных науках.

Предложенная модель призвана повысить качество управленческих решений, принимаемых на основании данных об оценках достигнутых научных результатов и прогнозировании новых научных достижений.

СПИСОК ЛИТЕРАТУРЫ

1. Сюнтюренко О.В. Финансирование фундаментальных исследований: концептуальный облик системы поддержки принятия решений с использованием методов наукометрии и анализа данных // Информатика и её применения. – 2018. – Т. 12, №. 1. – С. 118-127.
2. Хорошевский В.Ф., Ефименко И.В. Семантические технологии в наукометрии: задачи, проблемы, решения и перспективы // Когнитивно-семиотические аспекты моделирования в гуманитарной сфере. – Казань: Изд-во Академии наук. – 2017. – С. 222-266.
3. Маркусова В., Котельникова Н., Золотова А., Шухаева А. Перспективные направления научных исследований: мировые и отечественные тенденции по БД SCI-E, 2009 и 2015 гг. // Информация и инновации. – 2017. – № S1. – С. 111-118.
4. Филатова Т.Е., Пономарёв В.В. Теория «черного лебедя» // Материалы XVI межвузовской научно-технической конференции «Новые технологии в учебном процессе и производстве». – Рязань: Индивидуальный предприниматель Жуков Виталий Юрьевич, 2018. – С. 486-488.
5. Hicks D., Wouters P., Waltman L., De Rijcke S., Rafols I. Bibliometrics: the Leiden Manifesto for research metrics // Nature. – 2015. – Vol. 520, № 7548. – P. 429-431.
6. Лазар М.Г. Плагиат в научных коммуникациях современной эпохи // Ученые записки Российского государственного гидрометеорологического университета. – 2019. – № 56. – С. 166-175.

7. Виноградова Т.В. Добросовестность в научных исследованиях: Аналит. обзор. – Москва: ИНИОН РАН, 2017. – 74 с.
8. Сидельников Ю.В. Системный анализ экспертного прогнозирования. – Москва: МАИ, 2007. – 453 с.
9. Методика расчета качественного показателя государственного задания «Комплексный балл публикационной результативности» для научных организаций, подведомственных Министерству науки и высшего образования Российской Федерации, на 2020 год . – 2020. – URL: https://minobrnauki.gov.ru/documents/?ELEMENT_ID=24754&sphrase_id=31318.
10. Виноградова Т.В. Библиометрия и социогуманитарные науки не совместимы? // Науковедческие исследования. – 2016. – № 2016. – С. 90-106.
11. Малашук Н.М., Павлова Е.А. Проблемы и методы оценки результативности научных исследований // Альманах научных работ молодых ученых Университета информационных технологий, механики и оптики. – 2017. – С. 173-177.
12. Фейгельман М.В., Цирлина Г.А. Библиометрический азарт как следствие отсутствия научной экспертизы // Управление большими системами: сборник трудов. – 2013. – №. 44. – С. 335-342.
13. Reia S.M., Fontanari J.F. Wisdom of crowds: much ado about nothing. – 2020. – URL: https://arxiv.org/pdf/2008.01485.pdf
14. Голубков Е.П. Методы принятия управленческих решений в 2-х ч. Часть 2 : учебник и практикум для вузов. – Москва: Изд-во Юрайт, 2020. – 249 с.
15. Михайлов О.В., Аристов И.В. «Гибридные» индексы Хирша в оценке научной деятельности // Науковедческие исследования. – 2015. – № 2015. – С. 110-116.
16. Tsai C.F., Hu Y.H., Ke S.W.G. A Borda count approach to combine subjective and objective based MIS journal rankings // Online Information Review. – 2014. – Vol. 38, № 4. – P. 469-483. DOI 10.1108/OIR-11-2013-0253.
17. Сакулин С.А., Алфимцев А.Н. К вопросу о практическом применении нечетких мер и интеграла Шоке // Вестник Московского государственного технического университета им. Н.Э. Баумана. Сер. "Приборостроение" – 2012. – №. 1(1). – С. 55-63.
18. Bornmann L., Hug S. Bibliometrics-based heuristics: What is their definition and how can they be studied? – Research note // El profesional de la información (EPI). – 2020. – Vol. 29. – №. 4. – e290420. DOI: 10.3145/epi.2020.jul.20.

Материал поступил в редакцию 15.01.21.

Сведения об авторе

КАЛАЧИХИН Павел Андреевич – кандидат экономических наук, ведущий научный сотрудник ВИНИТИ РАН, Москва
e-mail: pakalachikhin@viniti.ru

ИНФОРМАЦИОННЫЙ ПОИСК

УДК 81’322.2’37

О.Л. Голицына, А.С. Гаврилкина

Об одном подходе к выделению имён сущностей и связей в задаче построения семантического поискового образа*

Представлены методы и средства выделения имён сущностей и связей на основе лексико-синтаксических шаблонов в рамках задачи семантического индексирования текстов документов. Содержание текста рассматривается как совокупность отражаемых триплетами элементарных фактов, включающих имена сущностей и отношений (имманентных, ситуативных и структурно-лингвистических). Для типизации ситуативных отношений используется таксономия отношений, в которой классы включают лингвистические конструкции; имманентные отношения формируются на основе сети понятий (тезауруса). Для идентификации свойств сущностей используется таксономия свойств и единиц измерения. Предложенный подход позволяет использовать в качестве поискового запроса имена сущностей, имена отношений, а также элементарные факты и составленные из них завершённые смысловые конструкции.

Ключевые слова: *семантический поиск, семантический поисковый образ, обработка текста, извлечение фактов, онтология*

DOI: 10.36535/0548-0027-2021-03-3

ВВЕДЕНИЕ

Семантический поиск в настоящее время ассоциируется с двумя классами задач.

Первый класс – отбор документов, отвечающих информационной потребности, выраженной запросом. Это хорошо известные механизмы: фактографический и тематический поиски, поиск аналогов, которые в совокупности нацелены на снятие неопределенности объекта/предмета поиска (лингвистической, семантической, прагматической).

Второй класс – комплексный аналитический (семантический) поиск, ориентированный на задачи синтеза нового знания (мониторинг проектов, выявление оснований и ограничений, анализ новизны, выявление потенциально возможных связей; оценка соответствия документации установленным критериям и нормативным требованиям и т.п.).

Отличие семантического поиска от традиционного (информационно-библиографического) состоит, в первую очередь, в представлении формы и содержания поискового запроса и результата поиска. Если поисковый образ запроса (и, соответственно, документа) при традиционном поиске обычно формулируется в виде последовательности ключевых слов, в общем случае не связанных общим контекстом употребления, то семантический поисковый образ, как запроса, так и документа, может включать не только ключевые слова, но и связи (отношения) между ними. При этом речь идет как об использовании имманентных отношений, наличие которых определяется посредством тезаурусов, так и ситуативных отношений, зависящих от лингвистической формы изложения конкретных фактов. Кроме того, решения задач, обеспечиваемые семантическим поиском (особенно задач второго класса), основываются на соотнесении и систематизации информации и характеризуются комплексностью результата и вариантностью его использования. Это позволяет определить семантический поиск как поиск не самих документов, а фрагментов, отвечающих в совокупности информационной потребности.

Качество и возможности поиска в значительной степени определяются индексированием массива документов, которое призвано обозначить смысл каждого документа. И если традиционно смысл представляется поисковым образом, включающим множество независимых ключевых слов, то при семантическом поиске поисковый образ – это совокупность связанных фактов, включающих имена сущностей и отношений.

* Работа выполнена при поддержке Министерства науки и высшего образования РФ (проект государственного задания № 0723-2020-0036)

В настоящей работе рассматривается методика построения семантического поискового образа путем преобразования текста документа в совокупность элементарных фактов-триплетов – пар имен сущностей, между которыми в отдельном предложении выделяется фрагмент текста, возможно представляющий семантическую связь. Такие связи в [1] называют «поверхностными», а подходы к их извлечению – «открытым извлечением информации» (Open Information Extraction).

Задача «открытого извлечения информации» была впервые сформулирована в [2] применительно к системе TextRunner, использующей для получения триплетов машинное обучение с частичным привлечением учителя. Подход TextRunner взят за основу при создании системы WOE [3], в которой удалось значительно повысить точность и полноту извлечения триплетов.

Для английского языка коллективом авторов был разработан ряд систем [4-9] (ReVerb, R2A2, ArgLearner, OLLIE, SRLIE, RelNoun, BONIE, OpenIE4, CALM), в которых для извлечения информации используются шаблоны из частей речи. В [4] представлен метод, при котором в качестве отношений рассматриваются глаголы и глагольные конструкции. При формировании отношений в [5] помимо глагольных связей привлекаются также связи между существительным и прилагательным (с учетом контекста). Кроме того, триплет уже сам может выступать в роли объекта и/или субъекта (аналогично реификации в RDF). В одну из методик, рассматриваемых в [8], включена возможность построения *n*-арных отношений. В [6] представлены ролевые связи, построение которых основано на разборе составных существительных. В [7] при формировании сущностей привлекаются правила извлечения числовых аргументов как единиц измерения. В [10] описаны многоязычные системы, такие как DepOE, ArgOE, LSOE и др.

Приведенные выше системы ориентированы преимущественно на обработку текстов на английском языке. Для разработки приложений, работающих с текстами на русском, создан ряд отечественных инструментальных систем: RCO Pattern Extractor [11], Alex [12], LSPL [13], DSTL [14], Томита-парсер [15]. Все эти инструментальные средства обработки текстов объединяет общий подход, основанный на распознавании и извлечении конкретных языковых конструкций. Для описания лингвистических свойств этих конструкций используется формальный язык, и само описание свойств осуществляется в форме специальных шаблонов и правил, с помощью которых происходит распознавание в тексте различных объектов и фактов. Совокупность шаблонов и правил формирует образец, настроенный на решение конкретной задачи, например, на выявление наименований юридических лиц, товаров, адресов; поиск информации о месте или дате рождения; выявление определений, понятий и т.п. Отметим, что принципиальным отличием этого подхода является то, что вначале создаются шаблоны, исходя из особенностей предметной области и решаемой задачи, после чего извлекаются факты. Однако, поскольку для информационного поиска характерно большое разнообразие форм представления смыслов и видов потребностей, описанный подход к построению семантического поискового образа не представляется полностью адекватным.

В настоящей работе предложен ориентированный на русский язык подход, использующий шаблоны для извлечения связей и формирования триплетов. Для унификации отношений применяется таксономия отношений, построенная на универсальных категориях.

НЕДОСТАТКИ КООРДИНАТНОГО ИНДЕКСИРОВАНИЯ

Для всех известных механизмов документального информационного поиска общим является сопоставление поискового образа документа и поискового образа запроса. При этом качество поиска определяется структурой этих образов и используемым критерием смыслового соответствия.

Основа информационного документального поиска – координатное индексирование – процесс, который заключается в формировании описания содержания документа в виде совокупности дескрипторов, выбираемых из заранее созданных словарей понятий либо из текста документа, и обозначающих основные понятия этого документа.

Метод координатного индексирования базируется на положении, что основное смысловое содержание документа и информационной потребности может быть с достаточной степенью точности и полноты выражено соответствующим списком ключевых слов[1], которые явно или в скрытом виде содержатся в тексте.

При так называемом «чистом» координатном индексировании ключевые слова в поисковых образах никак не связаны. В простейшем случае документ считается соответствующим информационному запросу (и подлежит выдаче), если в поисковом образе этого документа содержатся все ключевые слова поискового предписания. При этом необходимо понимать, что сами поисковые механизмы не имеют средств «угадывания» смысла (например, принадлежность определенной предметной или проблемной области) или интерпретации термина, используемого в качестве ключевого слова. Отсутствие возможности исключить влияние синонимии, полисемии и омонимии естественного языка, а также выразить ситуативные и имманентные связи между реальными объектами, процессами и т.п., представленными на вербальном уровне в тексте, существенно снижает семантическую силу информационно-поисковых языков, основанных на координатном индексировании.

К недостаткам «чистого» координатного индексирования [16], в основном, относят ложную или неполную координацию, когда в запросе недостаточно

[1] Под ключевыми словами в этом случае понимаются наиболее существенные для этой цели слова и словосочетания, обладающие назывной (номинативной) функцией. Назывные слова не обозначают предмет, а выделяют его путем указаний. К категории назывных слов относятся также имена собственные. Кроме назывных в качестве ключевых слов могут выступать и соответствующие численные характеристики, хронологические данные, диапазоны температур, давления и т. д.

использовать только координатную связь между ключевыми словами (например, результат запроса «Поставщики баз данных» содержит документы, представляющие базы данных поставщиков), а также отсутствие возможностей выражения в запросе парадигматических и синтагматических связей между ключевыми словами (например, запрос «Продажа лука», не доопределенный контекстом, приведет к формированию результата, содержащего и документы, представляющие лук как растение, и документы, в которых лук – это оружие).

Частично устранить отдельные недостатки могут, например, технологии расширения запроса (с использованием тезаурусов, лингво-процессоров, статистических связей и т.п.), однако они существенно зависят от особенностей понятийно-знаковой системы предметной области. Но, главное, их использование в автоматическом режиме на практике скорее ухудшает интегральные показатели эффективности поиска.

Для существенного повышения качества информационного поиска, основанного на применении координатного индексирования, необходимо разработать синтаксис информационно-поискового языка, который бы позволял использовать при построении поисковых образов документов и запросов не только простую координацию дескрипторов, но и существенные парадигматические и синтагматические связи.

СЕМАНТИЧЕСКИЙ ПОИСКОВЫЙ ОБРАЗ ДОКУМЕНТА

Поисковый образ линейной структуры, формируемый в процессе координатного индексирования, уже не отвечает задаче фиксирования не только ключевых слов, но и связей (отношений) между ними. В [17] предложен онтологический подход к построению поискового образа, который позволяет представить семантику документа системой понятий и отношений. Тогда при поиске в качестве запроса можно будет использовать завершенные смысловые конструкции.

Онтология определяется с позиций Общей теории систем как совокупность трёх взаимосвязанных систем: $O =< S_f, S_c, S_t \equiv>$ – функциональной (S_f), понятийной (S_c), терминологической (S_t) – и операции сопоставления элементов различных систем на уровне знаков (≡), обеспечивающей их тождество.

Функциональная система представляет объекты и отношения действительности средствами знакового уровня. Эти отношения имеют функциональную окраску, так как определяют способы и характер совместного существования и использования объектов. Логико-семантическим базисом онтологии является понятийная система, объектами которой служат устойчивые понятия предметной области, а набор отношений ограничен родовидовыми и ассоциативными (фиксируется в форме тезаурусов, рубрикаторов, классификационных схем и т.п.). Терминологическая система в онтологии отражает свойства естественного языка на уровне знаков – терминов, которые могут быть связаны отношениями эквивалентности (синонимии) и включения (образования словосочетаний). В качестве термина выступает отдельное слово или словосочетание естественного языка (или искусственного, например, шифр классификации), которое может применяться для описания понятия или объекта. Наличие понятийной и терминологической систем позволяет использовать для уточнения-расширения поискового запроса парадигматические отношения, формировать словосочетания, с разной степенью точности отражающие смысл.

В качестве моделей предлагаемых систем онтологии применяются помеченные ориентированные графы. При этом типологии вершин и дуг для графов понятийной и терминологической систем зафиксированы и однозначно заданы.

Семантической основой (смысловым «атомом») формирования графа функциональной системы является понятие элементарного факта – образа, фиксирующего некоторое состояние отдельного взаимодействия пары сущностей, где в роли сущности выступает понятие, объект, субъект и т.п., а взаимодействие представлено ситуативной связью (отношением). Элементарному факту в графе онтологии соответствует триплет «сущность – отношение – сущность», а множество вершин и множество дуг графа в совокупности соответствуют множеству элементарных фактов.

Таким образом, формирование онтологии, как поискового образа с сетевой организацией, требует:

1) задать понятийную систему онтологии с множеством парадигматических отношений;

2) представить текст документа в виде совокупности элементарных фактов.

Выражение имен сущностей и отношений на знаковом уровне позволяет индексировать элементарный факт как последовательность знаков, в которой представлены не только имена, но и типы сущностей и отношений. Таким образом могут быть построены как традиционные индексы (по ключевым словам), так и индексы, представляющие семантические связи. Наличие таких индексов позволит средствами традиционной теоретико-множественной модели информационного поиска, а также традиционного дескрипторного информационно-поискового языка реализовать отбор документов уже с учетом имманентных и ситуативных отношений между сущностями.

МЕТОДИКА ФОРМИРОВАНИЯ ЭЛЕМЕНТАРНОГО ФАКТА

Отображение смысла документа на множество элементарных фактов, формирующих узлы и дуги графа функциональной системы онтологии, основывается на классической схеме семантического анализа текстов, которая традиционно включает этапы графематического, морфологического, семантико-синтаксического и концептуального анализа [18].

На этапе графематического анализа выделяются структурные элементы текста (разделы, главы, абзацы, заголовки), текст разбивается на токены, которые идентифицируются и, в случае необходимости, объединяются с помощью словарей и лингвистических правил. Далее выявляются именные группы, даты, числа с плавающей точкой, аббревиатуры, единицы измерения и определяются границы предложений по знакам препинания с учётом идентифицированных специфических последовательностей символов. При

этом используются разделители элементов данных, разделители токенов, правила идентификации дат и аббревиатур, словари наименований и единиц измерения, разделители предложений.

Более точная идентификация токенов достигается за счёт учёта контекста их употребления, для чего используется таксономия свойств и единиц измерения [19]. Семантически значимые элементы текста, извлекаемые в границах одного или нескольких предложений, образующих семантическую окрестность токена, соотносятся с соответствующими компонентами онтологии, что позволяет восстанавливать недостающие смысловые фрагменты или выявлять противоречия, в частности, находить расхождения в обозначениях.

На этапе морфологического анализа происходит определение основных морфологических характеристик (часть речи, род, число, падеж) токенов, идентифицированных как слова, с использованием лингвистического процессора.

Этап семантико-синтаксического анализа начинается со снятия морфологической неоднозначности, порожденной на этапе морфологического анализа. Осуществляется выбор единственной парадигмы слова на основании анализа контекстного окружения и применения правил. После согласования морфологических характеристик каждое предложение преобразуется во множество триплетов – элементарных фактов.

В основе методики формирования элементарных фактов лежит представление отдельного предложения в виде линейной последовательности токенов, разделенной на синтаксические отрезки, идентифицируемые в соответствии с типологией, где тип задается множеством частей речи, к которым могут относиться токены отрезка. На начальном уровне каждый отрезок типизируется как «имя субъекта/объекта», «связь (часть связи)» или «разделитель». На следующем уровне отрезок типа «имя субъекта/объекта» доопределяется подтипами: «группа имени сущности», «группа подлежащего», «имя собственное», «аббревиатура», «значение и единица измерения». Отрезок типа «связь (часть связи)» типизируется как «действие» или «обстоятельство» и далее тип «действие» – подтипами «действие-глагол», «действие-причастие», «действие-краткое причастие», а тип «обстоятельство» – подтипами «предлог» и «контекст».

Для кодирования отрезков применяется система кодирования с единичной длиной кода, алфавит которой содержит заглавные буквы латинского алфавита и символ «#», кодирующий разделитель. В результате кодирования предложения формируется символьная строка, в которой выполняется поиск на соответствие лексико-синтаксическим шаблонам. Пример разбиения и кодирования фрагмента текста представлен на рисунке.

«Для площадки Курской АЭС-2 выполнен анализ возможных сценариев аварийных ситуаций, приводящих к возникновению воздушной ударной волны (ВУВ) от источников взрывной опасности, находящихся внутри площадки. Определены безопасные расстояния от внутриплощадочных источников возникновения ВУВ с давлением во фронте ВУВ, не превышающим 30 кПа по ПиН АЭ-5.6».

Для	площадки Курской АЭС-2	выполнен
F	N	K

анализ возможных сценариев аварийных ситуаций	,
S	#

приводящих	к	возникновению воздушной ударной волны	(	ВУВ	)	от
W	F	N	#	X	#	F

источников взрывной опасности	,	находящихся	внутри	площадки	.
N	#	W	F	N	#

Определены	безопасные расстояния	от
K	S	F

внутриплощадочных источников возникновения ВУВ	с	давлением	во	фронте ВУВ	,
N	F	N	F	N	#

не превышающим	30 кПа	по	ПиН АЭ-5.6	.
W	M	F	X	#

S - группа подлежащего; N - группа имени сущности; X – аббревиатура;
M - величина и единица измерения; K - действие-краткое причастие; W - действие-причастие; F – предлог; # - разделитель.

Пример разбиения и кодирования фрагмента текста

Токены «30» (распознан как числовое значение) и «кПа» объединены в результате проверки токена «кПа» на принадлежность к единицам измерения в таксономии свойств и единиц измерения. Для токена «кПа» однозначно определено свойство с наименованием «Давление»[2].

Для описания шаблонов разработан язык, позволяющий для триплета вида

<субъект(S)><связь(L)><объект(O)>

указать последовательности отрезков предложения, которые должны определять каждый компонент триплета. Шаблон состоит из двух частей, разделенных символом «=»: в левой части приводится подстрока для поиска в строке, кодирующей отдельное предложение, а в правой – задается порядок формирования триплета (элементарного факта) в следующем формате:

S<десятичная цифра>**L**{<десятичная цифра>}[…n(7)]**O**<десятичная цифра>

Десятичная цифра указывает на позицию символа в левой части шаблона. В соответствии с форматом после символа **L** может быть указано несколько десятичных цифр (не более 7-ми), каждая из которых определяет позицию отдельной части связи (т.е. связь может быть составной и не обязательно представляется одним непрерывным отрезком). При формировании подстроки можно задавать наборы символов в квадратных скобках, которые позволяют указать, что на данном месте в исходной строке может стоять последовательность из перечисленных символов произвольной длины. Например, простейший шаблон SVN=S1L2N3 дает возможность выявить в предложении элементарный факт, представленный триплетом <подлежащее><сказуемое><дополнение>. В зависимости от вида и жанра обрабатываемых документов могут быть сформированы разные наборы шаблонов.

Пример формирования элементарных фактов на материале приведенного нами фрагмента текста представлен в табл. 1 (буква «U» обозначает имя сущности, уже использованное в каком-либо триплете).

Таким образом, результатом этапа семантико-синтаксического анализа является формализация линейного текста до уровня совокупности триплетов, формирующих узлы и дуги графа функциональной системы онтологии. При этом связи в триплетах отражают выявленные в тексте ситуативные отношения между сущностями.

Формирование имманентных и структурно-лингвистических отношений, а также типизация ситуативных отношений, – задача, которая решается уже на этапе концептуального анализа.

Имманентные отношения на уровне функциональной системы выделяются для обеспечения входа в понятийную систему. Проводится проверка имен сущностей элементарного факта на равенство или «подобие» терминам понятийной системы (например, тезауруса). «Подобие» в этом контексте означает наличие среди отдельных слов имени всех слов, составляющих термин понятийной системы, в произвольном порядке (например, имя «система информационного поиска» будет эквивалентно термину тезауруса «информационно-поисковая система») или наличие всех слов термина среди составляющих имен обеих сущностей триплета (например, триплет S«язык»-L«поддерживать»-O«информационный поиск» будет соответствовать термину тезауруса «информационно-поисковый язык»). В случае обнаружения равенства или подобия термин понятийной системы включается во множество имен сущностей и формирует элементарный факт (триплет) с отношением «термин понятийной системы». Термин в этом типе отношения выступает в роли объекта, а субъектом становится имя сущности, равное или подобное термину (в случае подобия на уровне элементарного факта – имя сущности-субъекта).

Таблица 1

Шаблоны и соответствующие элементарные факты

Триплет	Шаблон
<анализ возможный сценарий аварийный ситуация> <выполнить для><площадка курский АЭС-2>	FNKS=S4L31O2
<источник взрывной опасность><находиться внутри><площадка>	N[#]WFN=S1L23O4
<анализ возможный сценарий аварийный ситуация> <приводить к><возникновение воздушный ударный волна>	S[#]WFN=S1L23O4
<возникновение воздушный ударный волна><контекст><ВУВ>	N[#]X= S1L0O2
<безопасный расстояние><определить от> <внутриплощадочный источник возникновение ВУВ>	KSFN=S2L13O4
<фронт ВУВ><не превышать><30 кПа>	N[#]WN=S1L2O3
<внутриплощадочный источник возникновение ВУВ><с><давление>	U[#]FN= S1L2O3
<30 кПа><по><ПиН АЭ-5.6>	U[#]FX= S1L2O3

[2] В случае, когда единице измерения соответствуют несколько наименований свойств из разных областей применения, для разрешения многозначности на этапе концептуального анализа может быть проведён поиск в тексте по связанным с идентифицированными свойствами компонентам таксономии (в пределах обрабатываемого предложения), или при помощи тезауруса проанализирована терминология в предложении и тексте в целом и уточнена область применения. Например, в рассматриваемом нами предложении присутствует термин «давление», который совпадает с названием свойства «Давление» в таксономии свойств и единиц измерения.

Таблица 2

Триплеты имманентных и структурно-лингвистических отношений

Имя сущности	Отношение	Имя сущности
анализ возможный сценарий аварийный ситуация	«Содержит сущность»	аварийный ситуация
аварийный ситуация	«Обстоятельство употребления»	анализ
аварийный ситуация	«Обстоятельство употребления»	возможный сценарий
площадка курский АЭС-2	«Содержит сущность»	АЭС-2
площадка курский АЭС-2	«Содержит сущность»	курский АЭС-2
курский АЭС-2	«Обстоятельство употребления»	площадка
источник взрывной опасность	«Содержит сущность»	взрывной опасность
взрывной опасность	«Обстоятельство употребления»	источник
возникновение воздушный ударный волна	«Содержит сущность»	воздушный ударный волна
воздушный ударный волна	«Обстоятельство употребления»	возникновение
внутриплощадочный источник возникновение ВУВ	«Содержит сущность»	ВУВ
фронт ВУВ	«Содержит сущность»	ВУВ
30 кПа	«Параметр»	ДАВЛЕНИЕ
Площадка	«Нижестоящий»	площадка курский АЭС-2
ВУВ	«Нижестоящий»	внутриплощадочный источник возникновение ВУВ
ВУВ	«Нижестоящий»	фронт ВУВ
анализ возможный сценарий аварийный ситуация	«Тезаурус»	АВАРИЙНЫЕ СИТУАЦИИ
возникновение воздушный ударный волна	«Тезаурус»	УДАРНЫЕ ВОЛНЫ

Структурно-лингвистические отношения формируются на основе распознавания аббревиатур внутри имени сущности, членения длинных словосочетаний, выявления имен сущностей – величин и единиц измерения, определения лексикографического включения имен сущностей.

При обнаружении аббревиатуры внутри имени сущности создается новый триплет со структурно-лингвистическим отношением: S<имя сущности> L<«Включает сущность»>O<аббревиатура>. Деление длинного словосочетания на два и более происходит в соответствии с правилами формирования словосочетаний тогда, когда оно содержит имя более чем одной сущности, например, при следовании прилагательного за существительным. В этом случае формируются новые триплеты со структурно-лингвистическими отношениями «Включает сущность» и «Обстоятельство употребления» (имени). Такие операции приводят к увеличению весового показателя коротких имен сущностей в тексте и позволяют выделить объекты-действия.

Имена сущностей (или части имен), идентифицированные на этапе графематического анализа как величины и единицы измерения, связываются отношением с наименованием измеряемого свойства в соответствии с таксономией свойств и единиц измерения [19].

В формировании структурно-лингвистических отношений лексикографического включения участвуют только имена сущностей элементарных фактов, построенных на этапе семантико-синтаксического анализа. Отношения строятся по принципу «от самого короткого имени к самому длинному» («Нижестоящий»), например, *нагрузка → пожарная нагрузка → постоянная пожарная нагрузка*.

В табл. 2 приведены триплеты имманентных и структурно-лингвистических отношений, построенные для содержимого табл. 1 на этапе концептуального анализа. В частности, имена АВАРИЙНЫЕ СИТУАЦИИ и УДАРНЫЕ ВОЛНЫ принадлежат тезаурусу, выполняющему роль понятийной системы онтологии.

Задача приведения различных естественно-языковых конструкций к единой модели на структурном уровне решается путем построения единой последовательности-тройки по шаблонам, представляющим различные последовательности текстовых единиц в предложении. Например, фрагментам текста «определены безопасные расстояния от внутриплощадочных источников возникновения ВУВ», «безопасные расстояния определены от внутриплощадочных источников возникновения ВУВ», «от внутриплощадочных источников возникновения ВУВ определены безопасные расстояния» будет соответствовать один триплет <безопасный расстояние><определить от> <внутриплощадочный источник возникновение ВУВ>, включающий ситуативное отношение «определить от». Однако такая структурная формализация не выявляет ситуаций, когда одна и та же семантика может быть реализована разными с точки зрения языкового выражения отношениями.

Таблица 3

Пример применения таксономии отношений

Отношение	Класс	Модальность
выполнить для	быть целью (предназначением)	достоверное| состоявшееся
находиться внутри	локативность в пространстве	достоверное| выполняющееся
приводить к	быть результатом	достоверное| выполняющееся
определить от	быть ограничением	достоверное| состоявшееся
не превышать	изменение	невозможное| выполняющееся
с	присоединение	
по	быть основанием	

Лексических конструкций, представляющих отношения в тексте, довольно много, и их значение зачастую доопределяется или изменяется контекстом их употребления.

Минимальной основой для построения исчисления семантики является иерархия классов сущностей, отношений и свойств, базирующаяся на общей системе характеристических признаков. Хотя обзор существующих решений показывает, что нет строгой и всеобщей классификации, тем не менее, есть ряд вполне самодостаточных решений [20].

Для типизации ситуативных отношений в рассматриваемой нами методике используется онтология отношений, основанная на трехуровневой иерархической классификации[3]: первый уровень отражает соотношение реальность/модель; второй – комбинации соотношений отдельного (часть) и агрегатного (целое); третий уровень построен по признаку формы проявления отношения – действие-ориентированные, объект-ориентированные и результат-ориентированные. Листьями иерархического дерева являются классы отношений, обладающие комбинацией свойств верхних уровней. Каждый класс содержит множество конкретных лингвистических конструкций, которые отражают его семантику. Классы нижнего уровня открыты для пополнения, и их содержимое может зависеть от вида текстового массива, подлежащего обработке. Для удобства восприятия и эксплуатации на основе онтологии путем линейно-иерархического упорядочения построена таксономия отношений, пример применения которой для типизации отношений, отраженных в табл. 1, представлен в табл. 3.

Отношения типа «Действие» дополнительно характеризуются модальными свойствами. Текущее состояние таксономии отношений позволяет выявить свойство, определяющее возможность осуществления действия со значениями «Достоверное (Актуальное)»/ «Предполагаемое» (возможное)/«Невозможное» и свойство, характеризующее состояние завершённости действия со значениями «Выполняющееся» / «Состоявшееся» / «Ожидаемое»[4].

Для установления значения первого модального свойства лингвистические конструкции отношений проверяются на наличие сигнальных слов и/или их сочетаний. Например, отношение «не превышать» в табл. 3 включает сигнальное слово «не», по которому определено значение модального свойства «Невозможное». Если сигнальные слова не обнаружены, устанавливается значение «Достоверное». Для определения значения второго модального свойства анализируются морфологические характеристики (время и вид) глаголов и отглагольных частей речи, входящих в отношение.

ИСПОЛЬЗОВАНИЕ ЭЛЕМЕНТАРНЫХ ФАКТОВ В ЗАДАЧАХ СЕМАНТИЧЕСКОГО ПОИСКА

В результате семантического поиска из фрагментов найденных по запросу документов должна быть построена новая единица знания. В этом контексте элементарный факт может рассматриваться как некий маркер конкретного смысла, содержащегося в отдельном предложении текста, – сохраненная при индексировании связь триплета с предложением дает возможность прямого перехода к изложению факта.

Такое использование элементарного факта существенно снижает требования к его семантической согласованности и завершенности – в качестве «проводника» к смыслу может рассматриваться только отдельный компонент (или пара компонентов). Например, решение задачи поиска определений понятий в отдельном тексте или в информационном массиве возможно свести к поиску предложений, в которых при индексировании был выявлен триплет, содержащий отношение из класса «Определение (понятия)». Отбор элементарных фактов по признаку принадлежности определенному классу отношений не ориентирован на формирование самого определения на базе триплета и, тем самым, не требует обязательного наличия корректного полного представления имен определяемого понятия и определяющих его сущностей. В табл. 4 приведены примеры фактов-триплетов и соответствующие им фрагменты текстов, отобранных по запросу на поиск определений.

[3] Более подробно см. в [20]

[4] Таким образом, всем отношениям типа «Действие» дополнительно присваиваются два значения модальности, однако возможно определение двух значений первого модального свойства для одного отношения.

Таблица 4

Фрагменты текста, соответствующие триплетам с отношением из класса «Определение (понятия)»

Триплет с отношением из класса «Определение (понятия)»	Фрагмент текста
<энергосистема планета> <*быть понимать под*> <энергетический система>	Под энергетической системой будем понимать энергосистему планеты либо континента, либо группы стран, либо страны, либо области страны, либо района области, либо в виде одной энергоплощадки, в которой размещены энергоблоки, или даже один энергоблок.
<текущий концепция конечный захоронение> <*являться так называть*> <Pollux-концепция>	Текущей концепцией конечного захоронения облученных топливных сборок является так называемая Pollux-концепция.
<пространство><<*называться*> <гермообъем>	Пространство, ограничиваемое защитной оболочкой, называется гермообъемом.
<параметр i> <*назвать*> <предел Высикайло>	Параметр *i* - безразмерное число, назовем его пределом Высикайло и равно оно с большой точностью 0,9•10-18 для любых квазинейтральных КДС Космоса, состоящих из любых химических элементов или веществ.
<способ> <*пониматься под*> <альтернативный источник энергия>	Соответственно, под альтернативным источником энергии понимается способ, устройство или сооружение, позволяющее получать электрическую энергию (или другой требуемый вид энергии) и заменяющее собой традиционные источники энергии, функционирующие на углеводородах.

Таблица 5

Фрагменты текста, соответствующие триплетам с отношением из класса «Результат (исследования)»

Триплет с отношением из класса «Результат (исследования)»	Фрагмент текста
<массив экспериментальный данные> <*быть получить в результат*> <испытание материал>	Массив экспериментальных данных, на основании которого установлены наши зависимости, был получен в результате испытаний материалов, которые подвергались облучению, в основном при температуре ~270°С.
<зависимость потеря транспортный свойство линия> <*не быть обнаружить в*> <эксперимент>	В экспериментах не было обнаружено зависимости потери транспортных свойств линии от наличия масляной пленки на поверхности электрода.
<расчетный анализ циклический прочность> <*быть провести применить к*> <контейнер>	Поскольку конструкции контейнеров, размещаемых на выгородке и на корпусе реактора ВВЭР-1000, принципиально не различаются, расчетный анализ циклической прочности был проведен применительно к контейнерам, устанавливаемым на выгородке, поскольку они подвергаются более жестким условиям нагружения.
<расчет упругий термический напряжение> <*провести в*> < работа>	В работе проведен расчет упругих термических напряжений, возникающих при работе ядерного реактора в таблетках ядерного топлива цилиндрической формы.
<использование код KESS> <*показать для*> <анализ процесс>	Показано использование кодов KESS, FREQN, FRADEMO, IDEMO и других для анализа процессов при авариях.

Однако следует отметить, что полнота и точность поиска во многом определяются составом и наполнением классов отношений. Развитие онтологии отношений, начиная с нижнего уровня классификационного дерева, происходит путем деления класса, сформированного на основе комбинаций значений свойств верхних уровней, с дальнейшим наполнением полученных подклассов конкретными лингвистическими конструкциями. При этом, если количество классов нижнего уровня ограничено произведением количества значений классификационных признаков первых трех уровней, то дальнейшее деление может происходить без привязки к системе признаков верхнего уровня. Это означает использование в рамках отдельного класса собственных признаков деления, возможное несбалансированное развитие отдельных типологий и даже нарушение принципа иерархии (возможность построения класса как объединения подклассов двух и более родительских классов). Значения признаков деления формируют семантику класса и определяют его конкретное наполнение. Такой способ построения онтологии позволяет иметь несколько видов ее существования: как универсальной, с ориентацией на предметную область, так и с ориентацией на множество решаемых задач. Иными словами, каждая онтология может предполагать свое лингвистическое наполнение.

В табл. 5 приведены примеры поиска триплетов и соответствующих им фрагментов текста по принадлежности отношений классу «Результат (исследования)» в научных статьях по атомной энергетике.

Формирование структурно-лингвистических отношений, типы которых предопределены, направлено на упорядочение лексики семантического поискового образа. С одной стороны, при выявлении связей типа «Включает сущность» и «Обстоятельство употребления (имени)» происходит деление имен сущностей в триплетах с целью выявления внутренних взаимосвязей: из длинных словосочетаний вычленяются более общие понятия, что приводит к увеличению частоты таких понятий и, соответственно, к возрастанию значения меры их семантической значимости в отдельном тексте или в документальном массиве в целом. С другой стороны, построение над именами сущностей лексикографических деревьев позволяет проследить тематическое развитие отдельного понятия от общего к более частному употреблению и использовать эти точные имена сущностей при поиске. Например, ветви лексикографического дерева понятия «электрод» содержат словосочетания:

- «никелевый электрод», «висмутовый электрод», «жидкометаллический электрод» и т.п., формируя родовидовые связи;
- «взрыв электрода», «нагрев электрода», «поляризация электрода» и т.п., отражая связи типа «объект-процесс»;
- «неприемлемое увеличение эффективного сопротивления электродов», «изучение поведения материала электродов МИТЛ», «нагрев электрода магнитоизолированной транспортирующей линии протекающим током» и т.д., указывающие на конкретные ситуативные факты.

Таким образом, можно утверждать о формировании при индексировании мини-тезаурусов или ситуативных рубрикаторов, которые способны служить для пользователя своеобразными когнитивными проводниками в задачах извлечения знания.

ЗАКЛЮЧЕНИЕ

Предложенная в настоящей работе методика формирования семантического поискового образа рассматривает его как часть трехуровневой онтологии, включающей множество элементарных фактов, множество точек входа в понятийную систему (тезаурус) и деревья лексикографического включения. Представляющие элементарные факты триплеты, извлекаемые из текстов способами, подобными изложенному в настоящей статье, хотя и отражают так называемые поверхностные связи, тем не менее довольно полно идентифицируют содержащийся в тексте смысл.

Использование таксономии отношений как дополнительного лингвистического обеспечения позволяет для конкретной лингвистической конструкции определить тип отношения и построить унифицированную (с точностью до типов отношений, включенных в таксономию) теоретико-графовую модель текста, и тем самым обеспечить сопоставимость смыслов, выраженных разными лингвистическими конструкциями. Индексирование текста как совокупности триплетов позволит в рамках традиционной теоретико-множественной модели информационного поиска (и средствами традиционного дескрипторного информационно-поискового языка) реализовать отбор документов уже с учетом имманентных и ситуативных отношений между сущностями. Приведенные здесь примеры показывают конструктивность применения предложенной методики в задачах семантического индексирования и поиска.

СПИСОК ЛИТЕРАТУРЫ

1. Шелманов А.О. и др. Открытое извлечение информации из текстов. Часть I. Постановка задачи и обзор методов // Искусственный интеллект и принятие решений. – 2018. – № 2. – С. 47.
2. Banko M. et al. Open information extraction from the web // Proceedings of the 20th International Joint Conference on Artifical Intelligence. – San Francisco: Morgan Kaufmann Publishers Inc., 2007. – P. 2670–2676/
3. Wu F., Weld D.S. Open information extraction using Wikipedia // Proceedings of the 48th annual meeting of the association for computational linguistics. – Uppsala: Association for Computational Linguistics, 2010. – P. 118-127.
4. Fader A., Soderland S., Etzioni O. Identifying relations for open information extraction // Proceedings of the 2011 conference on empirical methods in natural language processing. – Edinburgh: Association for Computational Linguistics, 2011. – P. 1535-1545.
5. Schmitz M. et al. Open language learning for information extraction // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Lan-

guage Learning. – Jeju Island: Association for Computational Linguistics, 2012. – P. 523-534.
6. Pal H. et al. Demonyms and compound relational nouns in nominal open IE // Proceedings of the 5th workshop on automated knowledge base construction. – San Diego: Association for Computational Linguistics, 2016. – P. 35-39.
7. Saha S. et al. Bootstrapping for Numerical Open IE // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers). – Vancouver: Association for Computational Linguistics, 2017. – P. 317-323.
8. Mausam M. Open information extraction systems and downstream applications //Proceedings of the twenty-fifth international joint conference on artificial intelligence. – Palo Alto: AAAI Press, 2016. – P. 4074-4077.
9. Saha S. et al. Open information extraction from conjunctive sentences // Proceedings of the 27th International Conference on Computational Linguistics. – Santa Fe: Association for Computational Linguistics, 2018. – P. 2288-2299.
10. Glauber R., Claro D.B. A systematic mapping study on open information extraction // Expert Systems with Applications. – 2018. – Vol. 112. – P. 372-387.
11. Ермаков А.Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor: компонент выделения особых объектов в тексте // Сб. трудов XII Международной научной конференции «Информатизация и информационная безопасность правоохранительных органов».– М.: Акад. упр. МВД России, 2003 – С. 312-317.
12. Жигалов В.А. и др. Система Alex как средство для многоцелевой автоматизированной обработки текстов // Компьютерная лингвистика и интеллектуальные технологии. – М.: ФГУП "Изд-во "Наука", 2002. – С. 192-208.
13. Большакова Е.И., Ефремова Н.Э., Шариков Г.Ф. Инструментальные средства для разработки систем извлечения информации из русскоязычных текстов // Новые информационные технологии в автоматизированных системах. – 2015. – № 18. – С. 533-543.
14. Скатов Д.С., Ливерко С.В., Окатьев В.В. Язык описания правил в системе лексического анализа ЕЯ-текстов Dictascope Tokenizer // Компьютерная лингвистика и интеллектуальные технологии: по материалам Международной конференции. «Диалог» (Бекасово, 26-30 мая 2010 г.). – 2010. – Т. 9, № 16. – С. 442-449.
15. Томита-парсер. Руководство разработчика. – URL: https://yandex.ru/dev/tomita/doc/dg/concept/about.html (дата обращения: 28.12.2020).
16. Михайлов А.М. Черный А.И., Гиляревский Р.С. Основы информатики. – М.: Наука, 1968. – 756 с.
17. Голицына О.Л., Максимов Н.В., Окропишина О.В., Строгонов В.И. Онтологический подход к идентификации информации в задачах документального поиска // Научно-техническая информация. Сер. 2. – 2012. – № 5. – С. 1-10; Golitsyna O.L., Maksimov N.V., Okropishina O.V., Strogonov V.I. The ontological approach to the identification of information in tasks of document retrieval // Automatic Documentation and Mathematical Linguistics. – 2012. – Vol. 46, № 3. – P. 125-132.
18. Белоногов Г.Г., Быстров И.И., Новоселов А.П., Козачук М.В., Хорошилов Ал-др А., Хорошилов Ал-сей А. Автоматический концептуальный анализ текстов // Научно-техническая информация. Сер. 2. – 2002. – № 10. – С. 26-32. Belonogov G.G., Bystrov I.I., Novoselov A.P., Kozachuk M.V., Khoroshilov A.A., Khoroshilov A.A. Automatic conceptual text analysis // Automatic Documentation And Mathematical Linguistics. – 2002. – Vol. 36, № 5. – P. 57-65.
19. Maksimov N. et al. Ontology of Properties and its Methods of Use: Properties and Unit extraction from texts // Procedia Computer Science. – 2020. – Vol. 169. – P. 70-75.
20. Максимов Н.В., Гаврилкина А.С., Андронова В.В., Тазиева И.А. Систематизация и идентификация семантических отношений в онтологиях научно-технических предметных областей // Научно-техническая информация. Сер. 2. – 2018. – № 11. – С. 32–42; Maksimov N.V., Gavrilkina A.S., Andronova V.V., Tazieva I.A. Systematization and identification of semantic relations in ontologies for scientific and technical subject areas // Automatic Documentation And Mathematical Linguistics. – 2018. – Vol. 52, № 6. – P. 306-317.

Материал поступил в редакцию 30.12.20.

Сведения об авторах

ГОЛИЦЫНА Ольга Леонидовна – доцент, кандидат технических наук, доцент института Интеллектуальных кибернетических систем Национального исследовательского ядерного университета «МИФИ», Москва.
e-mail: olgolitsina@yandex.ru

ГАВРИЛКИНА Анастасия Сергеевна – аспирант Национального исследовательского ядерного университета «МИФИ», Москва.
e–mail: asgavrilkina@yandex.ru

СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК [004.9:81’32’33]:025.3

А.Б. Антопольский

Цифровые лингвистические информационные ресурсы. Определение объекта и каталогизация

Обсуждается типология лингвистических информационных ресурсов (ЛИР), ставших важным инструментом прикладной лингвистики и информатики. Предлагается аналитический обзор международных организаций и проектов, специализирующихся в области ЛИР. Приводятся перечни зарубежных и российских каталогов, архивов и репозиториев ЛИР. Для развития ЛИР подчеркивается перспективность платформы связанных лингвистических открытых данных.

Ключевые слова: *лингвистические информационные ресурсы, типология, каталоги, архивы, репозитории, коллаборация*

DOI: 10.36535/0548-0027-2021-03-4

ВВЕДЕНИЕ

В последние десятилетия важнейшим инструментом как научных исследований в области теоретического и прикладного языкознания, так и практики применения компьютерных технологий в индустрии обработки текста и устной речи (далее – NLP[1]) стали цифровые (электронные) лингвистические информационные ресурсы (далее – ЛИР).

ЛИР – это наиболее актуальный и быстро развивающийся объект и инструмент цифровой гуманитаристики, один из существенных результатов научной деятельности в лингвистике, а также важное средство обучения языкам.

Доступность и качество ЛИР в значительной степени определяют эффективность коммуникации ученых и практиков, а также коллабораций в трудоемких процессах извлечения и представления лингвистической информации.

В России, как и во всем мире, создается большое количество лингвистических информационных ресурсов, и, как следствие, возникает множество организаций и проектов, ставящих цели координации, каталогизации, обмена, взаимного использования этих ресурсов и других методов их оптимизации. В настоящей статье дается краткий обзор этих институций и подходов к систематизации и каталогизации ЛИР.

ОБЪЕКТ ИССЛЕДОВАНИЯ И ТИПОЛОГИЯ ЛИНГВИСТИЧЕСКИХ ИНФОРМАЦИОННЫХ РЕСУРСОВ

При попытке систематизации и каталогизации ЛИР центральным является вопрос типологии, т.е. какие типы информационных объектов следует отнести к этой категории. Составители каталогов, справочных систем и репозиториев ЛИР придерживаются существенно различных взглядов на эту проблему.

Такие типы ЛИР как корпуса и лексиконы, безусловно, принадлежат к этой категории. Однако многие другие типы ЛИР не всеми составителями включаются в каталоги и справочные системы. Например, это:

- традиционные виды научной продукции лингвистики в электронной форме – публикации, отчеты, диссертации, труды конференций, библиографии;
- программные средства для NLP;
- географические информационные системы лингвистических данных;
- энциклопедические данные;
- популярные и учебно-образовательные ресурсы по изучению языков;
- памятники письменности;
- материалы диалектологических и этнографических экспедиций;
- лингвистическая документация по описанию редких языков;
- лингвистические институции (научные, образовательные, общественные, просветительские и др.)

[1] NLP – Natural Language Processing – обработка естественного языка

Так, Навигатор информационных ресурсов по языкознанию (http://niryaz2.alexo.beget.tech/), разработанный при нашем участии, в отличие от большинства каталогов ЛИР, содержит описания не только цифровых, но и бумажных ЛИР, в частности, библиотечных фондов, архивных и музейных документов. В настоящей статье будут рассматриваться только цифровые ЛИР.

Англоязычная Википедия указывает: «По состоянию на май 2020 года широко используемая стандартная типология языковых ресурсов не была создана (текущие предложения включают LRE Map, META-SHARE и, для данных, классификацию LLOD). Важные классы языковых ресурсов включают:

1) данные:

- лексические ресурсы, например, машиночитаемые словари ,
- лингвистические корпуса, т. е. цифровые коллекции данных на естественном языке,
- лингвистические базы данных, такие как коллекция кросс-лингвистических связанных данных;

2) инструменты:

- лингвистические аннотации и инструменты для создания таких аннотаций в ручном или полуавтоматическом режиме (например, инструменты для аннотирования подстрочного сглаженного текста, такие как Toolbox и FLEx , или другие инструменты языковой документации),
- приложения для поиска и извлечения таких данных (системы управления корпусом), для автоматического аннотирования (тегирование части речи , синтаксический анализ , семантический анализ и т. д.);

3) метаданные и словари:

- словари, репозитории лингвистической терминологии и языковых метаданных, например, META-SHARE (для метаданных языковых ресурсов), реестр категорий данных ISO 12620 (для лингвистических функций, структур данных и аннотаций в языковом ресурсе) или база данных Glottolog (идентификаторы для языковых разновидностей) и библиографическая база данных)»[2].

Приведем описания упомянутых нами типологий.

В карте оценочного описания ЛИР (LRE map http://lremap.elra.info/), разработанной в сообществе ELRA (Европейская ассоциация лингвистических ресурсов), выделено 3 основных типа: данные, документация и инструменты.

Список видов для типа ***ЛИР-данные*** выглядит следующим образом:

- Корпус
- Лексикон
- Онтология
- Грамматика / Языковая модель
- Терминология
- Банки деревьев зависимостей

ЛИР-документация:

- Данные для оценки ЛИР
- Инструменты оценки ЛИР
- Оценочные пакеты ЛИР
- Руководства и стандарты для оценки ЛИР
- Руководства для представлении и аннотаций ЛИР
- Технологическая инфраструктура ЛИР
- Метаданные

ЛИР-инструменты:

- Инструменты для аннотаций
- Корпусные инструменты
- Распознавание именованных сущностей
- Инструменты машинного перевода
- Программный инструментарий
- Токенизаторы
- Инструменты для машинного обучения
- Инструменты языкового моделирования
- Распознаватели неоднозначностей
- Распознаватели речи/ Транскриберы
- Обработка сигналов/ Извлечение свойств
- Веб-сервисы
- Синтезаторы текста в речь
- Идентификаторы языка
- Распознаватели говорящего
- Инструменты анализа настроений
- Просодические анализаторы
- Анализаторы изображений
- Инструмент устного диалога

META-SHARE (http://www.meta-share.org/) – это открытая сеть репозиториев для обмена языковыми данными, инструментами и связанными с ними веб-сервисами. В ней используется следующая типология:

- Корпуса
- Лексические концептуальные ресурсы
- Инструменты и сервисы
- Описания языков.

Типология LLOD (Linguistic Linked Open Data http://linguistic-lod.org/ – подробно о LLOD см. далее)

- Корпуса
- Лексиконы и словари
- Терминологические ЛИР, тезаурусы, базы знаний
- Метаданные ЛИР
- Категории лингвистических данных
- Типологические базы данных
- Другие.

Приведем еще несколько известных типологий ЛИР, наиболее полная из которых представлена в популярном ресурсе LINGUIST List (https://old.linguistlist.org/about.cfm). Основные разделы этого ресурса выглядят так:

- Люди и организации
- Вакансии
- Конференции и другие мероприятия
- Публикации
- Языковые ресурсы
  - Словари
  - Языки
- Области лингвистики
- Лингвистические компьютерные средства

Типология ЛИР авторитетной европейской лингвистической сети CLARIN (https://www.clarin.eu/), напротив, включает только ЛИР в узком смысле – это:

Корпуса

- компьютерных сетей
- научных текстов
- исторические

[2] Языковой ресурс - https://ru.qaz.wiki/wiki/Language_resource

- учебных текстов
- литературные
- аннотированные вручную
- мультимедийные
- газетные
- параллельные
- парламентские
- справочные
- устной речи

Лексические ресурсы
- лексика
- словари
- концептуальные ресурсы
- глоссарии
- списки слов

Инструменты
- нормализация
- распознавание именованных сущностей
- маркировка и лемматизация частей речи
- инструменты для анализа эмоционального восприятия.

Другая известная лингвистическая сеть OLAC (Консорциум открытых лингвистических архивов)[3] использует систему метаданных Дублинское ядро (DC). При этом для типов ресурсов DC предлагается расширение, включающая всего три квалификатора: словари, первичные тексты, лингвистическая документация.

В рамках сообщества ELRA создана также служба идентификации ЛИР – Международный стандартный номер ЛИР (ISLRN)[4], где используется типология ЛИР, принятая в OLAC.

Всё это подтверждает отсутствие единого подхода к определению объема понятия лингвистические информационные ресурсы, хотя пересечения всех их типологий легко видны.

МЕЖДУНАРОДНЫЕ ОРГАНИЗАЦИИ В СФЕРЕ ЛИНГВИСТИЧЕСКИХ ИНФОРМАЦИОННЫХ РЕСУРСОВ

CLARIN (Common European Research Infrastructure for Language Resources and Technology – https://www.clarin.eu/) – "Общие языковые ресурсы и технологическая инфраструктура" в Европе получила наибольшую известность. Эта инфраструктура была создана в 2012 г. для того, чтобы все цифровые языковые ресурсы и инструменты по всей Европы и за ее пределами были доступны через единую онлайн-среду для поддержки исследователей в области гуманитарных и социальных наук. Руководящий орган CLARIN – ERIC (European Research Infrastructures Consortium) – европейский консорциум исследовательской инфраструктуры.

Большинство операций, услуг и центров инфраструктуры CLARIN финансируются членами CLARIN ERIC (и наблюдателями), которыми могут быть страны или межправительственные организации. В странах созданы Национальные консорциумы, обычно состоящие из университетов, научно-исследовательских институтов, библиотек и государственных архивов, из которых по крайней мере один, имеет статус CLARIN-центра Ожидаемый вклад от членов и наблюдателей – создание и обеспечение доступа к цифровым языковым коллекциям данных, а также к цифровым инструментам и экспертным знаниям для работы с ними исследователей. В настоящее время в CLARIN входит 20 стран-членов, 3 страны-наблюдателя и еще одна организация.

ELRA (European Language Resources Association – http://www.elra.info/en/) – Европейская ассоциация лингвистических ресурсов – это другая общеевропейская структура в данной области. Основанная в 1995 г., ELRA является некоммерческой организацией, основная миссия которой заключается в том, чтобы сделать ЛИР, применяемые для технологий естественного языка, доступными для сообщества в целом. Для достижения этой цели ELRA осуществляет широкий спектр мероприятий вокруг ЛИР, включая идентификацию и распространение, производство и валидацию, оценку технологий, распространение информации о ЛИР.

В рамках ELRA действует несколько функциональных структур, в том числе, службы: Международного стандартного номера лингвистических ресурсов (International Standard Language Resource Number (ISLRN) – http://www.elra.info/en/islrn/), обеспечивающая идентификацию ЛИР; обмена ЛИР META-NET, функционирующая на основе специального механизма META-SHARE (http://www.elra.info/en/catalogues/meta-share/); каталогизации ЛИР; юридической поддержки ЛИР (Legal Support Helpdesk http://www.elra.info/en/services-around-lrs/legal-support-helpdesk/); управления данными (Data Management Plan http://www.elra.info/en/services-around-lrs/dmp/). Для реализации коммерческих задач при ELRA создано Агентство по оценке и дистрибуции ЛИР – ELDA (The Evaluations and Language resources Distribution Agency – http://www.elra.info/en/about/elda/).

ELEXIS (European lexicographic infrastructure – https://elex.is/) – Европейская лексикографической инфраструктура специализируется на разработке и дистрибуции инструментов для лексикографической деятельности; предоставляет бесплатный доступ к инструментам и инфраструктуре, разработанным партнерами проекта для академических институтов ЕС. Услугами ELEXIS пользуются более 400 организаций ЕС. Перечень инструментов и ресурсов ELEXIS размещен в Интернете (Tools and services – https://elex.is/tools-and-services/).

ELRC (European Language Resource Coordination https://lr-coordination.eu/**)** – Европейская служба координации лингвистических ресурсов для поддержки многоязычия в Европе действует при поддержке Connecting Europe Facility (CEF). Это фонд Европейского Союза для общеевропейских инфраструктурных инвестиций в транспортные, энергетические и цифровые проекты, направленные на расширение связей между государствами – членами Европейского Союза. ELRC поддерживает ряд ЛИР, а также услуг стран – членов ЕС в сфере языковых технологий.

TELRI (Trans-European Language Resources Infrastructure – http://telri.nytud.hu/**)** – Трансъевропейская инфраструктура лингвистических ресурсов – общеевропейский альянс, состоящий из 28 основных на-

[3] OLAC Metadata. – URL: http://www.language-archives.org/OLAC/metadata.html

[4] International Standard Language Resource Number (ISLRN) http://www.elra.info/en/islrn/

циональных языковых (технологических) учреждений с акцентом на страны Центральной и Восточной Европы и СНГ. В настоящее время функционирует 2-я очередь под названием TELRI II. От России в её состав входит Институт русского языка РАН.

OLAC (Open Language Archives Community http://www.language-archives.org/**)** – Сообщество открытых лингвистических архивов – наиболее известная в мире организация в сфере ЛИР, основана в 2000 г., представляет собой международное партнерство учреждений и отдельных лиц, которые создают всемирную виртуальную библиотеку языковых ресурсов путем: (1) выработки консенсуса относительно наилучшей современной практики цифрового архивирования языковых ресурсов и (2) развития сети взаимодействующих хранилищ и служб для обеспечения хранения и доступа к таким ресурсам. OLAC включает 62 архива, в общей сложности содержащих свыше 300 тыс. лингвистических информационных ресурсов.

Работу OLAC поддерживают Следующие службы:

- регистрация архива – сервис для проверки и регистрации архивов OLAC;
- автономные метаданные – сервис для проверки и форматирования записи метаданных OLAC;
- агрегатор OLAC-сервис, предоставляющий репозиторий, содержащий записи из всех других зарегистрированных репозиториев OLAC, включающий переход OLAC-DC и функцию запроса;
- Viser – это виртуальный сервис, позволяющий сайтам языковых ресурсов размещать сервисы на основе метаданных OLAC без необходимости сбора метаданных или разработки пользовательского интерфейса.

LDC (Linguistic Data Consortium https://www.ldc.upenn.edu/**)** – Консорциум лингвистических данных, созданный в Университете Пенсильвании. Открытый консорциум университетов, библиотек, корпораций и правительственных исследовательских лабораторий был образован в 1992 г. для решения критической проблемы нехватки данных, с которой в то время сталкивались исследования и разработки в области языковых технологий.

Первоначально основная роль LDC заключалась в хранении и распространении языковых ресурсов. На сегодня с помощью своих членов LDC превратилась в организацию, которая создает и распространяет широкий спектр языковых ресурсов, а также поддерживают спонсируемые исследовательские программы и оценки языковых технологий, предоставляя ресурсы и внося свой вклад в разработку языковых технологий. Кроме поддержки главного каталога ЛИР, консорциум ведет ряд проектов в области индустрии обработки языка.

LINGUIST List (https://old.linguistlist.org/ about.cfm) – Международное лингвистическое онлайн сообщество – наиболее популярная общественная организация, специализирующаяся на ЛИР. На портале LINGUIST List представлено множество ресурсов и сервисов, связанных с лингвистической деятельностью, как академической, так и прикладной. Организация поддерживается Департаментом лингвистики Университета Индиана (США), на её портале размещены справочные данные о разнообразных информационных объектах, связанных с лингвистикой:

- Люди и организации
- Вакансии и работа
- Мероприятия и конференции
- Публикации
- Лингвистические ресурсы
- Инструменты и программные средства
- Образовательные и учебные средства
- Поисковые средства
- Интерактивные сервисы
- Списки для рассылки.

TEI (Text Encoding Initiative https://tei-c.org/**)** – Инициатива по кодированию текста. Академическое сообщество, заинтересованное в практике семантической разметки текстов, в настоящее время поддерживает одноименный технический стандарт, собрания и серии конференций, а также журнал, вики, репозиторий GitHub, список рассылки и набор инструментов. Руководящие принципы TEI, давно считающиеся стандартом де-факто при подготовке цифровых текстовых ресурсов в научном исследовательском сообществе, предлагают огромный спектр потенциальных применений для кодирования текста – от традиционных научных изданий до языковых корпусов, исторических словарей, цифровых архивов и за их пределами. Анализ возможностей TEI для корпусной лингвистики подробно рассматривается В.П. Захаровым[5].

ISCA (International Speech Communication Association https://www.isca-speech.org/iscaweb/index.php/about-isca**)** – Международная ассоциация речевого общения осуществляет – содействие в международном мировом контексте деятельности и обменам во всех областях, связанных с наукой и технологиями речевого общения. Ассоциация предназначена для всех лиц и учреждений, заинтересованных в фундаментальных исследованиях и технологическом развитии, цель которых – описание, объяснение и воспроизведение различных аспектов человеческого общения с помощью речи. Прежде всего, это фонетика, лингвистика, компьютерная речь, распознавание и синтез, компрессия речи, распознавание говорящего, средства медицинской диагностики патологий голоса.

Список основных международных организаций, специализирующихся в области ЛИР, компьютерной лингвистики, а также цифровой гуманитаристики приводится в *Приложении 1*. Обширный список лингвистических обществ и ассоциаций имеется по адресу https://old.linguistlist.org/sp/GetWRListings.cfm?WRAbbrev=Societies.

МЕЖДУНАРОДНЫЕ ПРОЕКТЫ В СФЕРЕ ЛИНГВИСТИЧЕСКИХ ИНФОРМАЦИОННЫХ РЕСУРСОВ

GOLD (The GOLD Community of Practice (GOLDComm) http://linguistics-ontology.org/**)** – цель проекта объединить ученых, заинтересованных в наилучшей практике кодирования лингвистических данных. Проект продвигает передовую практику, развивает интероперабельность данных за счет использования

[5] Захаров В.П. Международные стандарты в области корпусной лингвистики // Структурная и прикладная лингвистика. – 2012. – № 9. – С. 201-221. ISSN 0202-2400.

стандарта GOLD, облегчает поиск по разрозненным наборам данных и предоставляет платформу для обмена существующими данными и инструментами. Стандарт GOLD охватывает лингвистические понятия, определения этих понятий и отношения между ними в свободно доступной онтологии.

LEGO (Lexicon Enhancement via the GOLD Ontology http://lego.linguistlist.org/**)** – цель проекта: создание инструментов и стандартов, облегчающих обмен и взаимодействие лексических данных с помощью онтологии GOLD; реализуется совместно с LINGUIST List (в настоящее время в Университете Индианы) и в Университете в Буффало.

ODIN (The Online Database of Interlinear Text http://odin.linguistlist.org/**)** – цель: сбор данных межлинейных текстовых примечаний (IGT – Interlinear Glossed Text) из Интернета, чтобы облегчить лингвистические исследования. В настоящее время интегрируется в рамках проекта GOLD ODIN с лингвистической онтологией, чтобы пользователи могли искать примечания, используя терминологию GOLD.

E-MELD (Endangered Metadata for Endangered Languages Data http://emeld.org/**)** – цель: создание архитектуры эффективного сотрудничества лингвистов, работающих над исчезающими языками; выработка консенсуса в отношении стандартов для метаданных, лингвистических аннотаций и языковой идентификации, что позволит обеспечить широкий доступ к данным в максимально полезной форме**.**

MultiTree (https://old.linguistlist.org/projects/multitree.cfm**)** – электронная библиотека научных гипотез о языковых отношениях и подгруппах, которые систематизируются в базе данных с возможностью поиска с помощью веб-интерфейса; каждая гипотеза представляется графически в виде интерактивного гиперболического отображения генеалогического древа, сопровождаемого информацией обо всех задействованных языках, а также об авторах и библиографических источниках гипотезы.

LL-MAP (Language and Location - A Map Annotation Project http://llmap.org/**)** – проект, предназначенный для интеграции языковой информации с данными физических и социальных наук с помощью географической информационной системы (ГИС). Онлайновый механизм сбора геоданных позволит ученым создавать карты на основе собственных наблюдений и дополнять полученную из открытых источников информацию о распределении языков.

Mailing Lists (https://old.linguistlist.org/lists/**)** – проект по созданию единого постоянного и доступного для поиска архивного сайта для сотен языковых форумов и дискуссий прошлого и настоящего, чтобы информация, которую они содержат, могла быть свободно доступна любому специалисту в этой дисциплине

RELISH (Rendering Endangered Language Lexicons Interoperable through Standards Harmonization https://old.linguistlist.org/projects/relish.cfm**)** – проект направлен на гармонизацию ключевых европейских и американских стандартов, установление единого способа представления структуры лексики в ЛИР, а также разработку процедуры миграции гетерогенных лексиконов в совместимый со стандартами формат XML.

MULTEXT-East (Multilingual Text Tools and Corpora for Central and Eastern European Languages http://nl.ijs.si/ME/**)** – многоязычные текстовые инструменты и корпусы для центрально- и восточноевропейских языков представляют собой многоязычный набор данных для лингвистических инженерных исследований и разработок.

Linport (The Language Interoperability Portfolio Project www.linport.org**)** – совместный проект по разработке открытого независимого от поставщика формата, который может быть использован многими переводческими сервисами для представления переводческих материалов.

Rosetta (https://rosettaproject.org/**)** – проект глобального сотрудничества языковых специалистов и носителей языка, работающих над созданием общедоступной цифровой библиотеки человеческих языков.

Проектов международных коллаборaций лингвистов известно гораздо больше. Их подробный список доступен, например, на портале LINGUIST List (https://old.linguistlist.org/sp/GetWRListings.cfm?WRAbbrev=Projects).

КАТАЛОГИ, АРХИВЫ И РЕПОЗИТОРИИ ЛИНГВИСТИЧЕСКИХ ИНФОРМАЦИОННЫХ РЕСУРСОВ

В Интернете имеется много различных собраний лингвистических информационных ресурсов, либо сведений о них. Эти собрания имеют форму порталов, содержащих ссылки на ЛИР, каталогов, поисковых и справочных систем, архивов и репозитариев, созданных по различным принципам, с разным количеством отраженных в них ЛИР и различными моделями их описания. Суммарное количество ЛИР, отраженных в этих собраниях, превышает 300 тыс. Всего нами обнаружено свыше 160 таких собраний, которые приводятся в *Приложении 2*. Значительная часть этих собраний включена в перечни архивов OLAC (Open Language Archives Community Participating Archives http://www.language-archives.org/archives), лингвистических мета-сайтов LINGUIST list (Language Meta Sites https://old.linguistlist.org/sp/GetWRListings.cfm?wrtypeid=25) или каталог репозиториев научных данных RE3 (Registry of research data repositories https://www.re3data.org/). Однако эти перечни существенно пересекаются, поэтому мы сочли полезным сделать общий список. При этом около 20 архивов и репозиториев центров ЛИР европейских стран входят в сеть CLARIN и построены по единым стандартам и методикам.

Особый тип ЛИР представляют терминологические базы и банки данных (ТБД). Их каталог можно найти, например, по адресу: Terminology websites & blogs https://termcoord.eu/terminology-websites. Специфика ТБД заключается в том, что их создают не столько лингвистические, сколько международные организации широкого профиля или отраслевые (ООН, ЕС, ISO. ФАО и др.), используются они в основном для переводческой и редакторской деятельности.

Российские каталоги лингвистических информационных ресурсов приводятся в *Приложении 3*. Архивов и репозиториев ЛИР в России нет. Отметим,

что в 2012 г. Д.А. Усталов опубликовал первый анализ российских каталогов ЛИР[6], причем в сферу его рассмотрения вошло всего 5 каталогов. Сейчас их значительно больше – в нашем списке их свыше 40, причем в этот список не вошли каталоги образовательных ресурсов по русскому языку. Таких каталогов очень много, их топ-10 в качестве примера приведены в *Приложении 4*.

Связанные лингвистические открытые данные

Наиболее перспективным способом координации ЛИР и развития эффективных коллабораций в этой области является, по нашему мнению, платформа Семантического веба и связанных открытых данных, которая позволяет реально обеспечить одноразовое создание и многократное использование данных, причем лингвистические данные для этой платформы подходят как нельзя лучше. Ученые, работающие на платформе Семантического веба, создали портал (Linguistic Linked Open Data http://linguistic-lod.org/), на котором размещено облако LLOD (Linguistic Linked Open Data – связанные лингвистические открытые данные). Данное облако уже сейчас включает более 200 ЛИР, и оно автоматически пополняется при вводе новых лингвистических информационных ресурсов.

ЗАКЛЮЧЕНИЕ

На основе исследования в области ЛИР по гранту РФФИ[7], мы предложили проект интеграции лингвистических информационных ресурсов, по крайней мере, создаваемых в учреждениях Российской академии наук на основе Центра ЛИР, который предлагается основать при Институте русского языка им. В.В. Виноградова РАН. Основные результаты этого исследования приводятся в монографии автора[8]. Как продолжение этого исследования предполагается разработать Автоматизированную справочную систему «Русский язык» на базе ИНИОН РАН и ИРЯ РАН, которая должна включать, как минимум, каталог, а в идеале – репозиторий ЛИР по русскому языку.

Краткий обзор международных институций и проектов в области лингвистических информационных ресурсов, а также их собраний, с очевидностью доказывает необходимость координации этой деятельности в России, в том числе, для предлагаемого проекта, и возможность широкого использования международного опыта**.**

[6] Усталов Д.А. Каталоги лингвистических ресурсов: состояние и перспективы // Молодой ученый. – 2012. – № 12(47). – С. 148-152. – URL: https://moluch.ru/archive/47/5955/ (дата обращения: 02.01.2021).

[7] Грант РФФИ 18-00 – 00298 КОМФИ «Интеграция научно-информационных ресурсов учреждений РАН по гуманитарным наукам (на примере языкознания) как части единого цифрового информационного пространства РАН».

[8] Антопольский А.Б. Научная информация и электронное пространство знаний : монография // Фундам.б-ка. / под науч. ред. Д.В. Ефременко. – Москва : ИНИОН, 2020. – 313 с. ISBN: 978-5-248-00964-0. DOI: 10.31249/spaknow/2020.00.00. – URL: http://inion.ru/ru/publishing/publications/nauchnaia-informatciia-i-elektronnoe-prostranstvo-znanii/

Приложение 1

Международные организации в области лингвистических информационных ресурсов

ACH Association for Computers and the Humanities Ассоциация компьютеров и гуманитарных наук https://ach.org/about-ach

ACL Association for Computational Linguistics Ассоциация компьютерной лингвистики https://www.aclweb.org/portal /

ADHO Alliance of Digital Humanities Organizations Альянс организаций цифровой гуманитаристики https://wiki2.org/en/Alliance_of_Digital_Humanities_Organizations

AFNLP Asian Federation of Natural Language Processing ST Азиатская федерация по обработке естественного языка http://www.afnlp.org/wp/

COCOSDA, International Committee for the Coordination & Standardisation of Speech Databases and Assessment Techniques. Международный Комитет по координации и стандартизации речевых баз данных и методов оценки http://www.cocosda.org/

CLARIN Common European Research Infrastructure for Language Resources and Technology https://www.clarin.eu/

EADH European Association for Digital Humanities Европейская ассоциация цифровой гуманитаристики https://eadh.org/

ELRA European Language Resources AssociationЕвропейская ассоциация лингвистических ресурсов http://www.elra.info/en/

Humanistica L'association francophone des humanités numériques/digitales http://www.humanisti.ca/

IAMT International Association for Machine Translation Международная ассоциация машинного перевода http://eamt.org/international-association-for-machine-translation/

ICCL (COLING) International Committee on Computational Linguistics Международный комитет количественной лингвистики https://wiki2.org/en/International_Committee_on_Computational_ Linguistics

ISCA International Speech Communication Association Международная ассоциация речевых коммуникаций https://www.isca-speech.org/iscaweb/index.php/about-isca

IQLA The International Quantitative Linguistics Association Международная ассоциация количественной лингвистики http://www.iqla.org/

Linguistic diversity and multilingualism on Internet. Программа ЮНЕСКО по поддержке языкового разнообразию и многоязычию в Интернете http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/linguistic-diversity-and-multilingualism-on-internet/

LTAC Language Terminology/Translation and Acquisition Consortium Консорциум терминологии и переводов http://ltacglobal.org/about.html

SIL International — международная некоммерческая организация, бывший Летний Институт Лингвистики (Summer Institute of Linguistics) https://www.sil.org/

TEI Text Encoding Initiative https://tei-c.org/

TC ISO 37 Технический комитет ИСО 37 Лингвистика и терминология
https://www.iso.org/ru/committee/48104.html
TerminOrgs. Terminology for Large Organizations Терминология для крупных организаций – консорциум терминологов / http://www.terminorgs.net

Приложение 2

Мировые каталоги, архивы и репозитарии лингвистических информационных ресурсов

A Digital Archive of Research Papers in Computational Linguistics http://www.ldc.upenn.edu/acl
Academia Sinica Collections
https://ndaip.sinica.edu.tw/en_3-1-2-7.html
AfBo (A world-wide survey of affix borrowing)
http://afbo.info//
AFLAT (African Language Technologies)
https://www.aflat.org/
African Language Materials Archive
http://www.aiys.org/aodl/public/access/alma_ebooks/
AILLA (Archive of the Indigenous Languages of Latin America) http://www.ailla.utexas.org/
Alaska Native Language Archive
http://www.uaf.edu/anla
ANPERSANA bibliothèque numérique
https://anpersana.iker.univ-pau.fr
APiCS Online http://apics-online.info//
ARCHE (A Resource Centre for the HumanitiEs)
https://arche.acdh.oeaw.ac.at/browser/
Arquivo.pt - o Arquivo da Web Portuguesa
http://www.arquivo.pt
ASEDA (Aboriginal Studies Electronic Data Archive)
http://www1.aiatsis.gov.au/ASEDA/ASEDAsr.xml
AusNC (Australian National Corpus)
https://www.ausnc.org.au/
BAS Repository (Bavarian Archive for Speech Signals, BAS CLARIN Repository, Bayerisches Archiv für Sprachsignale) https://clarin.phonetik.uni-muenchen.de/BASRepository
BathSPAdata (Bath Spa University figshare)
https://bathspa.figshare.com/
Best Language Websites https://sites.uni.edu/becker/
BowPed TRPS Data http://repository.edition-topoi.org/collection/TRPS
Buckeye Speech Corpus http://buckeyecorpus.osu.edu/
Burckhardt Source (The European correspondence to Jacob Burckhardt) https://burckhardtsource.org/
BVH (Bibliothèques Virtuelles Humanistes, Humanistic Virtual Libraries) http://www.bvh.univ-tours.fr/
California Language Archive http://cla.berkeley.edu
CEDIFOR (Repository CLARIN-D Centre)
https://www.cedifor.de/repository-clarin-d-centre-cedifor-2
C'ek'aedi Hwnax Ahtna (Regional Linguistic and Ethnographic Archive) http://hdl.handle.net/10125/4538
CELR META-SHARE (Center of Estonian Language Resources, Eesti Keeleressursside Keskus)
https://keeleressursid.ee/en/resources
Central Institute of Indian Languages: Publications
http://www.ciil.org/Main/Publications/publication.asp
CHILDES Data repository http://childes.talkbank.org/
CLAPOP (The Dutch CLARIN Portal Pages)
http://portal.clarin.nl
CLARIN Center BBAW (CLARIN service center of the Zentrum Sprache at the BBAW) https://clarin.bbaw.de
CLARIN Center Tübingen (CLARIN repository at the University of Tübingen)
https://uni-tuebingen.de/en/ 134314
CLARIN Centre Vienna (CCV, Language Resources Portal, LRP, CLARIN-AT) https://clarin.oeaw.ac.at/ccv/
CLARIN INT Center (CLARIN IvdNT-portaal)
https://portal.clarin.inl.nl/
CLARIN Resource Families
https://www.clarin.eu/resource-families
CLARIN.SI (Slovenian language resource repository)
http://www.clarin.si/
CLARIN: el inventory of language resources and services https://inventory.clarin.gr/
CLARIN-DK-UCPH Repository (The CLARIN Centre at the University of Copenhagen)
https://repository.clarin.dk/repository/xmlui/
CLARIN-ERIC (Common Language Resources and Technology Infrastructure – European Research Infrastructure Consortium) https://www.clarin.eu/
CLARIN-LT http://clarin-lt.lt/?lang=en
CLARINO Bergen Center repository
https://repo.clarino.uib.no/xmlui/
CLARIN-PL (Language Technology Centre, Centrum Technologii Jezykowiych) https://clarin-pl.eu/dspace/
CLARIN-UK https://www.clarin.ac.uk/
CNC (Czech National Corpus, CNK, Český národní korpus) https://korpus.cz/
CNRTL (Centre National de Ressources Textuelles et Lexicales) https://www.cnrtl.fr/
CoCoON (COllections de COrpus Oraux Numeriques, CoCoON ex-CRDO)
http://cocoon.huma-num.fr/ exist/crdo/
Codex Sinaiticus Experience the oldest Bible
http://www.codexsinaiticus.org/en/
Comparative Corpus of Spoken Portuguese
http://www.ime.usp.br/~tycho/
Dades DD Dipòsit Digital (Dipòsit Digital de la Universitat de Barcelona)
http://diposit.ub.edu/dspace/handle/2445/56364
DAIS (Digital Archive of the Serbian Academy of Sciences and Arts, Digitalni arhiv izdanja SANU)
https://dais.sanu.ac.rs
DaSCH (Data and Service Center for the Humanities)
http://data.dasch.swiss
DeReKo (Das Deutsche Referenzkorpus, Das Portal für die Korpusrecherche in Textkorpora des Instituts für Deutsche Sprache, The Mannheim German Reference Corpus)
https://www1.ids-mannheim.de/kl/projekte/ korpora/
DGD (Datenbank Gesprochenes Deutsch, DGD2 (formerly), FDZ AGD, Forschungsdatenzentrum Archiv für Gesprochenes Deutsch am Institut für Deutsche Sprache Database for Spoken German)
https://dgd.ids-mannheim.de/
Dictionaria http://dictionaria.clld.org//
D-PLACE (Database of Places, Language, Culture and Environment) https://d-place.org

DSAL (Digital South Asia Library)
http://dsal.uchicago.edu/
DTA (Deutsches Textarchiv)
http://www.deutschestextarchiv.de/
ELP (English Lexicon Project)
https://elexicon.wustl.edu/
ELRA Catalogue of Language Resources
http://catalogue.elra.info/
Endangered Languages Archive
https://lat1.lis.soas.ac.uk/ds/asv/
Ethnologue: Languages of the World
http://www.ethnologue.com
Eurac Research CLARIN Centre http://clarin.eurac.edu/
Forenames_histHub
https://thesaurus.histhub.ch/forenames/en
GAMS (Humanities' Asset Management System, Geisteswissenschaftliches Asset Management System)
http://gams.uni-graz.at/context:gams
GerManC project (A representative historical corpus of German 1650-1800)
https://www.alc.manchester.ac.uk/modern-languages/research/german-studies/germanc/
Glottolog 4.3 http://glottolog.org//
Graduate Institute of Applied Linguistics Library
http://www.gial.edu/library
HDC (Humanities Data Centre)
http://humanities-data-centre.de/
http://www.nb.no/English/Collection-and-Services/Spraakbanken
HunCLARIN http://clarin.hu/content/hunclarin-tagjai
Huygens ING https://www.huygens.knaw.nl/?lang=en
Huygens ING: eLaborate
http://elaborate.huygens.knaw.nl/
HZSK Repository (Hamburger Zentrum für Sprachkorpora Korpus Repositorium, Hamburg Centre for Speech Corpora Digital Repository)
https://corpora.uni-hamburg.de/hzsk/en/repository-search
IANUS (Datenportal Digitale Forschungsdaten aus Archäologie&Altertumswissenschaften)
https://www.ianus-fdz.de/datenportal/
IDR (Informatics Research Data Repository)
https://www.nii.ac.jp/dsc/idr/en/index.html
IDS Repository (IDS-Mannheim Repository, Institut für Deutsche Sprache Repository)
http://repos.ids-mannheim.de/
ILC-CNR for CLARIN-IT repository
http://www.clarin-it.it/
Ilovelanguages
http://www.ilovelanguages.com/index.php
IMS Universität Stuttgart Repository (IMS Fedora Repository, Repository of the CLARIN-D Centre at IMS Stuttgart) http://clarin04.ims.uni-stuttgart.de/repo
IULA UPF OAI Archive
http://iula02v.upf.edu/corpus_data/oai-iula/oai.pl
Kaipuleohone
http://scholarspace.manoa.hawaii.edu/handle/10125/4250/
Kielipankki (The Language Bank of Finland)
https://www.kielipankki.fi/language-bank/
LAC (Language Archive Cologne, KA³ Language Archive Cologne) https://lac.uni-koeln.de
Language Commons Language Corpora
http://www.archive.org/details/LanguageCommons
Language Documentation and Conservation
http://scholarspace.manoa.hawaii.edu/handle/10125/310/
Language Resource Inventory LINDAT/CLARIAH-CZ
https://lindat.cz/
Language resources at the Text Laboratory
http://www.hf.uio.no/tekstlab/
LAPSyD (Lyon-Albuquerque Phonological Systems Database) http://www.lapsyd.ddl.cnrs.fr/lapsyd/
LAUDATIO Long-term Access and Usage of Deeply Annotated Information https://www.laudatio-repository. org /
LDC (Linguistic Data Consortium)
https://www.ldc.upenn.edu/
Leipzig Corpora Collection)
http://clarin.informatik.uni-leipzig.de/repo/
LINGUIST List https://old.linguistlist.org/about.cfm
Linguistic Linked Open Data http://linguistic-lod.org/
Linguistics, Natural Language, and Computational Linguistics Meta-index
https://nlp.stanford.edu/links/linguistics.html
List of resources by language
https://aclweb.org/aclwiki/List_of_resources_by_language
Living Archive of Aboriginal Languages
http://laal.cdu.edu.au/
Lund University Humanities Lab corpusserver
https://corpora.humlab.lu.se
Magoria Books' Carib and Romani Archive
http://archive.magoriabooks.com/
Meertens Instituut Collecties (Meertens Institute Collections, De Digitale Koepel)
http://www.meertens.knaw.nl/cms/en/
Melbourne.figshare.com University of Melbourne data repository https://melbourne.figshare.com/
META-SHARE http://metashare.elda.org/
MICASE (Michigan Corpus of Academic Spoken English)
http://quod.lib.umich.edu/m/micase/
MMSH Phonothèque (Maison méditerranéenne des sciences de l'homme, Phonothèque, Mediterranean Research Centre for the Humanities)
http://phonotheque.mmsh.huma-num.fr/
MULCE (Multimodal Learning and teaching Corpora Exchange LETEC, MUltimodal contextualized Learner Corpus Exchange) http://lrl-diffusion.univ-bpclermont.fr/mulce2/accesCorpus/accesCorpusMulce.php
ODIN (The Online Database of Interlinear Text)
http://www.csufresno.edu/odin
OLAC Coverage http://www.language-archives.org/documents/ coverage.html
OLAC. Participating Archives
http://www.language-archives.org/archives
ORDO (Open Research Data Online, The Open University Data Repository) https://ou.figshare.com/
ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue, Open Resources and TOols for LanGuage) https://www.ortolang.fr/
OTA (The University of Oxford Text Archive)
http://ota.oucs.ox.ac.uk/
Pacific Collection at the University of Hawai'i at Mānoa Hamilton Library
http://library.manoa.hawaii.edu/departments/hp/pacific/about.php

PARADISEC (Pacific And Regional Archive for Digital Sources in Endangered Cultures) http://catalog.paradisec.org.au
PELIC The University of Pittsburgh English Language Institute Corpus https://github.com/ELI-Data-Mining-Group/PELIC-dataset
PhA (Phonogrammarchiv) https://www.oeaw.ac.at/phonogrammarchiv/
PHOIBLE 2.0 http://phoible.org//
POLLEX-Online http://pollex.org.nz
PolMine Project https://polmine.github.io/
PORTULAN CLARIN repository https://portulanclarin.net/
QMU eData Repository (Queen Margaret University eData Repository) https://eresearch.qmu.ac.uk/handle/20.500.12289/4
Repository CLARIN-D Centre Leipzig (CLARIN-D repository at the University of Leipzig,
RMS (Romani Morpho-Syntax Database) https://romani.humanities.manchester.ac.uk//rms/
SADiLaR (South African Centre for Digital Language Resources) https://www.sadilar.org/
SAILS Online http://sails.clld.org//
SIL Language and Culture Archives http://www.sil.org/resources/language-culture-archives
SinMin https://osf.io/a5quv/
SLDR Speech and Language Data Repository http://www.sldr.org
Spanish CLARIN Knowledge Centre http://www.clarin-es-lab.org/index-es.html
Sprachatlas Baden-Württemberg https://escience-center.uni-tuebingen.de/escience/sprachatlas/#8/48.676/8.992
Språkbanken (The Swedish Language Bank) https://spraakbanken.gu.se/eng
Språkbanken (Speech & Language Data Bank, Språkbankens ressurskatalog)
St. Edward's University institutional repository, St. Edward’s University Figshare https://stedwards.figshare.com/
Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources https://nlp.stanford.edu/links/statnlp.html
SWE-CLARIN https://sweclarin.se/eng/home
SWELANG (CLARIN Knowledge Centre for the Languages of Sweden, Clarin kunskapscentrum) http://www.sprakochfolkminnen.se/om-oss/forskning/sprakbanken-sam/clarin-kunskapscentrum/swelang.html
TALKBANK Data repository http://talkbank.org/
Tekstlaboratoriet (tekstlab, The Text Laboratory) http://www.hf.uio.no/iln/english/about/organization/text-laboratory/
Terminology websites & blogs https://termcoord.eu/terminology-websites
TextGrid Repository (Virtuelle Forschungsumgebung für die Geisteswissenschaften Virtual research environment for the Humanities) https://www.textgridrep.org/
The Crúbadán Project http://crubadan.org/
The Eclectic Company Language & Linguistics http://www-personal.umich.edu/~jlawler/lingmarks.html
The Language Archive https://archive.mpi.nl/oai2
The LDC Corpus Catalog http://catalog.ldc.upenn.edu/
The LINGUIST List Language Resources http://linguistlist.org/olac/
The Natural Language Software Registry http://registry.dfki.de
The Polinsky Language Sciences Lab Dataverse https://dataverse.harvard.edu/dataverse/polinsky
The Rosetta Project (A Long Now Foundation Library of Human Language) http://www.rosettaproject.org/
The Sociolinguistic Archive and Analysis Project (SLAAP) https://slaap.chass.ncsu.edu/
The speech-language resources http://www.speechlanguage-resources.com/
The Typological Database Project https://portal.clarin.nl/node/1920
Tibetan and Himalayan Digital Library http://iris.lib.virginia.edu/tibet/
TLA (The Language Archive) https://tla.mpi.nl/
TRACTOR https://www.slideserve.com/Gabriel/tractor
Transnewguinea – database of the languages of New Guinea http://transnewguinea.org/
TROLLing (Tromsø Repository of Language and Linguistics) https://trolling.uit.no
TST-Centrale (De Centrale voor Taal- en Spraaktechnologie or TST-Centrale) http://www.tst-centrale.org
U Bielefeld Language Archive http://www.spectrum.uni-bielefeld.de/langdoc/
UCLA Phonetics Lab Archive http://archive.phonetics.ucla.edu/
UdS Fedora Commons Repositorium http://fedora.clarin-d.uni-saarland.de/index.en.html
UK RED UK Reading Experience Database http://www.open.ac.uk/Arts/reading/UK/
UniLang Online https://unilang.org/resources.php?mode=category&sid=201b9efa0b49c56f98eeaee9d09ea730
University of Guelph Dataverse (University of Guelph Research Data Repository Dataverse) https://dataverse.scholarsportal.info/dataverse/ugrdr
WALS Online (World Atlas of Language Structures Online) http://wals.info/refdb_oai/
WALS Online RefDB http://wals.info/
Webonary Sites https://www.webonary.org
WOLD (The World Loanword Database) http://wold.clld.org//

Приложение 3

Российские каталоги лингвистических информационных ресурсов

Lexilogos. Русские словари для онлайн перевода https://www.lexilogos.com/english/russian_dictionary.htm
Linguists Ресурсы для переводчиков и лингвистов http://linguists.narod.ru/catalogue.html
NLPub — каталог ресурсов для обработки естественного языка. https://nlpub.ru/
Архив петербургской русистики www.ruthenia.ru/apr/index.htm
Ассоциация лингвистов-экспертов Юга России http://www.ling-expert.ru/links.html

Веб-сайты филологической и лингвистической тематики http://it.lang-study.com/veb-sajty-filologicheskoj-i-lingvisticheskoj-tematiki/
Все о языках, лингвистике, переводе... http://linguistic.ru
Информационные ресурсы по лингвистике http://homepages.tversu.ru/%7Eips/InfoSeek.htm
Каталог «Наука в Рунете» Лингвистика. https://elementy.ru/catalog/t123/Lingvistika
Каталог интернет-ресурсов по РКИ http://www.russischlehrer.at/fileadmin/Veranstaltungen/internetquellen_rki.pdf
Каталог лингвистических программ и ресурсов в Сети https://rvb.ru/soft/catalogue/index.html
Каталог лингвистических программ и ресурсов в Сети http://rykov-ling.narod.ru/resurs.html
Каталог лингвистических программ и ресурсов в Сети http://ru-writer.blogspot.com/2010/06/c.html
Каталог лингвистических программ и ресурсов в Сети. Программы анализа и лингвистической обработки текстов https://helpiks.org/1-109079.html
Компьютерная лингвистика https://elementy.ru/catalog?type=38
Компьютерная лингвистика. Портал знаний https://uniserv.iis.nsk.su/cl/
Лингвистика http://filologia.su/lingvistika/
Лингвистика в России: ресурсы для исследователей https://archive.vn/9Fu4X#selection-183.2-185.27
Лингвистические корпуса и сервисы http://web-corpora.net/
Лингвистические ресурсы https://www.sites.google.com/site/lingvistika586/lingvisticeskie-resursy
Лингвистические интернет-ресурсы https://didacts.ru/termin/lingvisticheskie-internet-resursy.html
Лингвистические ресурсы Интернета https://poisk-ru.ru/s71285t1.html
Лингвистический энциклопедический словарь http://tapemark.narod.ru/les /
Многоязычные сайты и универсальные списки ссылок http://linguodiversity.narod.ru/Links/multlang.htm
Навигатор информационных ресурсов по языкознанию http://niryaz2.alexo.beget.tech/
Общая филология (интернет-ресурсы) http://yspu.org/
Общие ресурсы по лингвистике и филологии http://www.garshin.ru/linguistics/linguistic-portals.html
Онлайн ресурсы для работы переводчика https://lingvadiary.ru/?p=199
Продукты Центра речевых технологий https://www.speechpro.ru/product/
РОССИЙСКАЯ ЛИНГВИСТИКА (RUSLING) http://rusling.narod.ru/index.htm
Русский язык http://www.philology.ru/linguistics2.htm
Русский язык https://russkiyyazik.ru /
Специализированные ресурсы по лингвистике https://studopedia.ru/13_129557_spetsializirovannie-resursi-po-lingvistike.html
Справочные интернет-ресурсы http://www.oshibok-net.ru/for-all/sites/210/
Тематические порталы, сайты → Языкознание http://library.altspu.ru/lang.phtml
Филологические ссылки. Лингвистика, языкознание http://konf-csu.narod.ru/ze/links.html
Филологический портал Philology.ru http://www.philology.ru/
Электронные ресурсы http://www.ling-theory.ru/links
Электронные ресурсы. Языкознание www.lib.tsu.ru
Энциклопедия русского языка https://ksana-k.ru/?p=1941
Языковые порталы и сайты http://library.mrsu.ru/content/tags/linguistics.html

Приложение 4

Электронные образовательные ресурсы по русскому языку (Top-10)

Единая коллекция цифровых образовательных ресурсов. Русский язык http://school-collection.edu.ru/catalog/search/?text=%D0%F3%F1%F1%EA%E8%E9+%FF%E7%FB%EA&submit=%CD%E0%E9%F2%E8&interface=catalog&subject%5B%5D=8
Мультимедийные интернет-ресурсы для изучения школьной программы по русскому языку и литературе https://xn--j1ahfl.xn--p1ai/library/multimedijnie_internetresursi_dlya_izucheniya_shkol_ 221352.html
Образовательные Интернет-ресурсы, направленные на поддержку и продвижение русского языка https://nsuem.ru/library/resources/rus-lang-resouces/
Перечень лицензионных электронных образовательных ресурсов по русскому языку и литературе https://multiurok.ru/files/perechen-litsenzionnykh-elektronnykh-obrazovatelny.html
Перечень электронных образовательных ресурсов для учителя русского языка и литературы https://infourok.ru/perechen-elektronnyh-obrazovatelnyh-resursov-dlya-uchitelya-russkogo-yazyka-i-literatury-4288134.html
Ресурсы по русскому языку http://www.den-za-dnem.ru/school.php?item=295
Электронные ресурсы открытого образования по русскому языку: Лучшие практики https://www.pushkin.institute/science/konferencii/eor_rus/Sbornik_ER.pdf
ЭОР для учителей русского языка и литературы https://nsportal.ru/shkola/literatura/library/2019/11/17/eor-dlya-uchiteley-russkogo-yazyka-i-literatury
ЭОР для учителя русского языка и литературы http://kykyshkina.semenovschool3-nn.edusite.ru/p11aa1.html
ЭОР по русскому языку https://rosuchebnik.ru/material/eor-po-russkomu-yazyku/

Материал поступил в редакцию 11.01.21.

Сведения об авторе

АНТОПОЛЬСКИЙ Александр Борисович – доктор технических наук, профессор, главный научный сотрудник ИНИОН РАН, Москва
e-mail: ale5695@yandex.ru