ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

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. **№ 9** **Москва 2017**

ОБЩИЙ РАЗДЕЛ

УДК 008

А.Д. Урсул

Информационный вектор эволюции: от цефализации к культурогенезу

Показано, что эволюционный прогресс происходит под влиянием и даже «давлением» информации и информационных процессов, в связи с чем можно констатировать наличие информационного вектора глобальной эволюции, и что с определённого этапа эволюция живого вещества идет в направлении цефализации, когда в биосистемах, наряду с генетической информацией, появилась новая информационная система, ответственная за экстракорпоральный «вынос» информационных процессов во внешнюю среду. Рассматриваются проблемы генетической связи процессов цефализации и культурогенеза, становления культуры как внегенетического и экзогенного информационного процесса. Отмечается, что возникновение тех или иных ограничений накоплению, передаче и переработке информации ведёт к формированию новых путей и способов преодоления этих границ и продолжению дальнейшего роста информационного содержания материальных систем на супермагистрали глобально-универсальной эволюции.

Ключевые слова: *глобальная эволюция, информация, информационная культурология, информационный вектор эволюции, культура, культурогенез, ноосфера, семиосфера, цефализация, экзогенного накопления информации принцип*

ВВЕДЕНИЕ

Информация и информационные процессы играют фундаментальную роль в эволюции вообще [1, 2], а особенно – в глобально-универсальной эволюции, представляющей собой своего рода «информационную стрелу» самоорганизации, пронизывающую всё мироздание. Этот тип эволюции является перманентной самоорганизацией материальных систем в неживой природе, живом веществе и обществе, продолжающейся далее в социоприродной форме и охватывающей вещественный фрагмент Вселенной [3-5].

Для оценки и даже количественного измерения степени развития уже несколько десятилетий используется способ определения изменения информационного содержания (и связанной с ним негэнтропии) материальных систем в ходе эволюционной самоорганизации (либо самодезорганизации). Причем, на прогрессивной линии эволюции происходит непрерывное накопление информации в системах, и тем самым это позволяет выявить информационный вектор развития материальных систем. Оказалось, что перманентная самоорганизация имеет место в основном в информационном аспекте как наиболее важном, «ответственном» за все эволюционные процессы. Именно информационное видение эволюции позволило сформулировать информационный принцип прогрессивной эволюции, которая реализуется лишь в случае накопления информации в развивающейся системе.

В ходе глобальной эволюции на главной её траектории с переходом от низшей ступени эволюционного ряда к высшей на пути усложнения материи увеличивается информационное содержание систем. Можно считать, что уже выявлен, достаточно обоснован и объективно существует упомянутый «информационный вектор» глобальной эволюции, получивший название супермагистрали, по которой идёт непрерывное накопление информации в развивающихся системах.

Вместе с тем, на супермагистрали глобальной эволюции, наряду с информационным критерием и вектором, реализуется тенденция сужения эволюционного коридора – сокращение пространства распространения и масс-энергетического объема возникающих все более сложных структур. С одной стороны, происходит рост информационного содержания в каждой более высокой структурной единице, увеличивается многообразие видов и форм существования все более высоких структурных уровней, а, с другой стороны, сужается их общий суммарный объем и пространство распространения в мироздании. Изменение этой тенденции сужения «эволюционного коридора» происходит лишь на социальной ступени эволюции, благодаря появлению культурно-экзогенного принципа накопления, переработки и других форм движения информации, что приводит к освоению все большего как планетарного, так и внеземного пространства.

Наиболее важную роль в «переломе» упомянутых тенденций развёртывания глобальной эволюции сыграли такие ещё мало изученные в информационном ракурсе процессы, как цефализация и культурогенез.

ЦЕФАЛИЗАЦИЯ КАК ИНФОРМАЦИОННЫЙ ПРОЦЕСС

Наиболее примечательной особенностью становления социальной ступени эволюции стал процесс цефализации. Именно феномен этот оказался наиболее важным для эволюционного перехода из биологического мира в мир социальный. Следуя принципам эволюционизма и элевационизма и прослеживая в прошлое траекторию антропосоциогенеза, можно обнаружить биологические истоки этого процесса в форме начала цефализационного процесса.

В 1944 г. в небольшой, но очень важной и своей последней статье «Несколько слов о ноосфере» В.И. Вернадский не просто обратил внимание на то, что эволюция живого вещества идет в направлении цефализации, но поставил эту идею по своему значению рядом с дарвинизмом [6, с. 235-244]. Как пишет В.И. Вернадский, это явление было названо Д.Д. Дана «цефализацией», а Ле-Контом «психозойской эрой». Д.Д. Дана, подобно Ч. Дарвину, пришел к этой мысли, к этому пониманию живой природы во время своего кругосветного путешествия, которое он начал через два года после возвращения в Лондон Ч. Дарвина, т. е. в 1838 г., и которое продолжалось до 1842 г.

«Дана указал, – отмечает В.И.Вернадский, – что в ходе геологического времени…, т. е. на протяжении двух миллиардов лет, по крайней мере, а наверное много больше, наблюдается (скачками) усовершенствование – рост – центральной нервной системы (мозга), начиная от ракообразных, на которых эмпирически и установил свой принцип Дана, и от моллюсков (головоногих) и кончая человеком. Это явление и названо им цефализацией. Раз достигнутый уровень мозга (центральной нервной системы) в достигнутой эволюции не идет уже вспять, только вперед» [6, с. 241].

Представляется, что в статье, посвященной становлению ноосферы и, по сути, завершающей мысли ученого по этой проблеме, вопрос о цефализации был поднят далеко не случайно. Уместно обратить внимание, что В.И. Вернадский вспомнил открытый ещё в 1851 г. Д.Д. Дана принцип цефализации, т.е. спустя почти сто лет после его открытия (всё это время научная общественность о нём попросту забыла). И важно раскрыть интуитивную догадку В.И. Вернадского, который, хотя и не прямо, но всё же указал на генетическую связь процесса цефализации и ноосферогенеза.

Кстати, и другой основоположник концепции ноосферы П. Тейяр де Шарден в широко известной книге «Феномен человека» [7] также особо выделял в эволюции мироздания процесс цефализации, на вершине которого оказался человек. Оговоримся, что между цефализацией и ноосферогенезом должен быть поставлен культурогенез, который стал необходимым этапом и продолжением информационного процесса цефализации и основой для развёртывания ноосферогенеза.

В словарях термин цефализация (от греч. *kephale* – голова) чаще всего трактуют как процесс дальнейшего усложнения организмов, обособления головы и включения в её состав органов, которые у предков были в иных местах организма; включения туловищ-

ных сегментов в головной отдел у животных в процессе их эволюции – на том конце тела животного, который обращен в сторону движения, т.е. вперед. Здесь также концентрируются органы чувств, передние отделы центральной нервной системы, управляющие работой этих органов и входящие в головной мозг, под контроль деятельности коры которого на высшей фазе цефализации попадает вся нервной система.

Цефализацию характеризуют и как увеличение отношения массы головного мозга к массе тела животного, что особенно видно в ряду позвоночных животных – от рыб до человека. Известно, что степень цефализации, как относительная масса мозга позвоночных, наиболее высока у птиц, а из млекопитающих – у приматов, особенно у человека, и китообразных. Цефализация в основном осознаётся как необратимая прогрессивная эволюция нервной системы животных, выражающаяся в сосредоточении нервных клеток на головном конце тела, которое в первую очередь встречается с внешней средой.

Однако это феноменологическое и фенотипическое описание не объясняет сути информационно-биологического процесса. Фенотип характеризуется как совокупность внешних и внутренних признаков организма, приобретённых в результате индивидуального развития, как «вынос» генетической информации в тело организма, навстречу факторам окружающей среды. Причем, некоторые признаки фенотипа напрямую определяются генотипом, в то время как другие зависят от взаимодействия организма с окружающей средой, в котором важнейшую роль играет получение и переработка информации. Тем самым роль живого вещества в эволюции планеты возрастала благодаря информационным факторам за счёт извлечения ими всё больших объёмов информации, всё более совершенной переработки этой информации и передачи её другим организмам [8, с. 23-25].

Фенотипические признаки лишь «внешне» выражают феномен цефализации между тем как в его основе лежат информационные процессы и, прежде всего, речь идёт об оперативной реакции и переработке поступающей из окружающей среды информации. Информация как научное понятие по историческим меркам только недавно вошла в «научный обиход» – всего немногим более полувека, но именно это понятие оказалось наиболее важным для понимания эволюции вообще и глобально-универсальной эволюции – в особенности.

«Двигателем» цефализационного процесса оказалась информация – в биосистемах, наряду с генетической информацией, появилась и стала развиваться новая весьма перспективная информационная система, которая оказала огромное влияние на весь дальнейший процесс биоэволюции. И не только этого этапа глобальной эволюции. Цефализация создала способ самоорганизации, который преодолел информационную ограниченность генетических биопроцессов, замкнувшихся в самом биологическом организме.

Если информация накапливается лишь в каком-либо организме, то дальнейшее её накопление не может происходить в самом организме без увеличения его размеров и других параметров. При этом появляются и другие ограничения и новые пределы дальнейшему росту информационного содержания биосистем, использующих информационно-генетический способ такого накопления. Поэтому необходима интеграция организмов (что произошло при переходе от одноклеточных к многоклеточным), либо накопление информации уже за пределами организма и его генетической системы. Цефализация оказалась такого рода «трамплином», открывшим путь развёртыванию нового способа накопления и переработки информации и нового этапа эволюции – социокультурной эволюции, причём важно, что «социальное» как «культурное» формировалось именно через информационные процессы.

Цефализация как перманентная эволюция центральной нервной системы позволяла осваивать всё более разнообразные формы взаимодействия организмов с окружающей средой. Появилась возможность наследования навыков, опыта жизнедеятельности, в результате чего усложнялась и генетическая информационная система живых организмов. Причём, параллельно происходила и цефализация, которая развивалась ускоренными темпами, и обогащение генетического кода. Это взаимодействие цефалических и генетических процессов и их взаимодействие с окружающей средой привело в итоге к появлению сознания и способности высших организмов формировать эндогенное накопление и способы переработки информации.

Именно благодаря цефализационному процессу и через него на стадии становления сознания человека произошел экстракорпоральный «вынос» информации во внешнюю среду, и это кардинально изменило весь процесс эволюции, трансформировав биологическую эволюцию в социальную, а точнее – в социокультурную эволюцию. Генетический процесс накопления и переработки информации в биосистемах постепенно, но во всё большей степени стал передавать эстафету процессу цефализации, с помощью которого удалось накапливать всё большее количество информации в отдельном организме и с помощью эстракорпоральных технологий перерабатывать её, передавая эту информацию в разных формах другим организмам.

Между прочим, уже отмеченное выше исчерпание одного из видов информационных биологических механизмов состоялось ранее – в ходе биоэволюции. В.А. Красилов, отмечая, что человек в эволюционном смысле уникален, так как его эволюция почти полностью смещена в область культуры, далее пишет: «В истории жизни на Земле ход эволюции дважды круто изменялся: первый раз на переходе от простейших к многоклеточным организмам, когда возможности биохимического совершенствования были в основном исчерпаны, прогресс сместился в сторону морфологии, и второй – в связи с возникновением человеческой культуры, принявшей эстафету прогресса от морфологии» [9, с. 89]. Кстати, заметим, что в обоих случаях накопление информации происходило за счёт внешней по отношению к тому или иному организму среды.

С этой точки зрения, культурогенез и его результат – культура предстает как новая форма дальнейшего постбиологического продолжения супермаги-

страли универсальной эволюции, которую уже не сдерживают биолого-генетические ограничения. Это наводит на мысль о том, что, если появляются какие-либо ограничения накоплению, передаче и переработке информации на супермагистрали глобально-универсальной эволюции, то находятся новые пути и способы преодоления этих границ и пределов, что демонстрирует переход как от одноклеточных к многоклеточным организмам, так и от генетического механизма накопления и движения информации к социокультурному способу.

Это подтверждает наличие в прошлом и продолжение в будущее «информационной стрелы» глобальной эволюции. Но, начиная с биологической ступени, эта стрела обретает форму своеобразного «информационного напора» самоорганизации как в ходе биоэволюции, так и в глобально-универсальной эволюции, поскольку поступательное движение вверх происходит под влиянием и даже своего рода «давлением» информации и информационных процессов. И это произошло потому, что эти информационные процессы уже на биологическом уровне обрели форму процессов управления, что существенно расширило возможности и горизонты овладения информацией окружающей среды. Тем более, что вышедшая из биологического мира социальная ступень эволюции трансформировала этот «информационный напор» в «информационный взрыв», благодаря новому – постгенетическому и эндогенно-культурному способу накопления и освоения информации.

ИНФОРМАЦИОННАЯ ПРИРОДА КУЛЬТУРОГЕНЕЗА

В ряде современных культурологических исследований культура представляется в качестве особой информационной системы, характеризующей сущность социальной ступени эволюции, в том числе и на её цивилизационной стадии. Культура появляется раньше перехода человечества к цивилизационному состоянию, поскольку появление культуры совпадает с самим появлением социальной ступени эволюции. Цивилизация же представляет собой уже достаточно поздний этап эволюции этой ступени, на которой находится современное человечество.

Причем, если понятие цивилизации «тяготеет» к материальной, вещественно-энергетической, экономической и технико-технологической трактовкам, представляя в то же время целостную социальную ступень эволюции, то понятие культуры акцентирует внимание, главным образом, на ее духовно-информационной составляющей, выражающей глубинную сущность социальной ступени эволюции. Материально-вещественные объекты цивилизации выступают как феномены культуры только в том случае, если они рассматриваются в информационном ракурсе, а именно – как системы такой информационной «суперсистемы», как семиосфера. Эти системы способны хранить и передавать созданные человеком знаки, имеющие не только ценность, но и, прежде всего, значение (смысл), и регулирующие деятельность человека и социума.

Среди множества концепций культуры и соответствующих им представлений культурогенеза появилась и символическая концепция культуры. Эта концепция была предложена ещё немецким философом и культурологом Э. Кассирером, который полагал, что сущностью культуры является символическая деятельность [10]. Человек, по Э. Кассиреру, это символическое животное, вышедшее из природного мира благодаря созданию искусственного мира символов, способности к систематической и массовой символизации, и сформировавшее коллективный интеллект и коллективную память, именуемую культурой. Символы и знаки являются основными видами социальной информации, благодаря которым культура кодирует и накапливает всё большее количество информации. Поэтому такой подход к интерпретации культуры считается информационно-семиотическим, представляя эту форму социального бытия как надиндивидуальный механизм хранения, накопления, переработки и других форм движения информации.

Понятий культуры достаточно много, как и вариантов культурогенеза, однако практически невозможно при определении или характеристике культуры не указать на её информационную составляющую. Очевидно, что культура представляет весьма сложный и многокомпонентный феномен, а это, в свою очередь, допускает и порождает многообразие дефиниций и значений, широких и узких.

Причём, к узким относится то понятие культуры, которое используется, например, в соответствующем ведомстве (министерстве) и вузах культуры нашей страны. «Культура» – это совокупность формальных и неформальных институтов, явлений и факторов, влияющих на сохранение, производство, трансляцию и распространение духовных ценностей (этических, эстетических, интеллектуальных, гражданских и т.д.)» – так формулируется определение этого понятия в принятых в 2014 г. «Основах государственной культурной политики» [11, с.7]. Как видим, даже это определение, несмотря на более узкий смысл трактовки, базируется на информационной основе, выделяя духовные ценности, формируемые человеческим сознанием, хотя и не содержит самого понятия информации.

Впрочем, в другом официальном документе, принятом уже в 2016 г., содержится даже некоторая критическая оценка этого подхода к ограничению сферы культуры: «Особенностью современного подхода к гуманитарной сфере является узковедомственный подход к культуре, а также отчасти утилитарное понимание культуры как сферы услуг. Это порождает более низкий общественный статус культуры…» [12, с. 18]. Из культуры как целостного социально-информационного феномена тем самым выделяется лишь её часть, в основном сфера духовной культуры.

В принципе необходимо исходить из того, что тот или иной аспект или элемент культуры есть везде – в любом социуме, во всей цивилизации, даже в сосуществующих в настоящее время племенах синполитейного палеолита. Трудно представить такую возможную ситуацию, чтобы культура когда-то возникла, выделив всю социальную ступень из биологической, а затем постепенно превратилась лишь в упомянутую выше сферу культурной деятельности. Иначе придётся полагать, что деятельность людей вне сферы такой сферы

как культура является «некультурной» и «несоциальной», что явно алогично и бессмысленно.

Поэтому, выделяя определённую сферу культуры как приоритетную область деятельности, следует иметь в виду, что «культурное» в той или иной степени существует во всех других социальных и цивилизационных процессах. Хотя ясно, что степень содержания «культурного» в этих процессах может быть различной, и это является важной, но пока малоисследованной проблемой как культурологии и социологии, так и цивилизационных исследований. Поэтому важно различать культуру как общечеловеческий феномен и ту узкую сферу деятельности, которой занимается так называемая «культурная деятельность».

Между тем, важно и нужно понять, почему именно информация выбирается в качестве фундаментальной основы интерпретации природы культуры. Ранее были попытки (например, американского культуролога Л. Уайта [13]) связать эволюцию культуры с изменением энергии, а уровень культурного развития определять по уровню энергопотребления. Но эти идеи не были в дальнейшем поддержаны исследователями, причём понятно почему: такого рода идеи просто далеки от природы феномена культуры.

И если на заре становления человечества основным источником энергии был организм человека, то впоследствии увеличение энергоресурсов развивающегося общества происходило за счет экстракорпоральных энергетических источников. Рост энергетического потенциала и могущества человечества так же, как и накопление информации, происходил за счет окружающей человека среды, и это вполне согласуется с принципами синергетики. Значит, дело не во внеорганизменном и внегенетическом накоплении информации, энергии и вещества, что в принципе характерно для человечества, а в накоплении информации, которая оказалась, по мнению многих исследователей, гораздо ближе к природе культуры, чем вещество и энергия. Хотя понятно, что без них культура, как информационный феномен, существовать также не может.

Для реализации внегенетических информационных функций уже сформировались цефализационные структуры, которые взяли на себя ответственность за вынос информационно-культурных процессов за пределы организма фомировавшегося древнего человека. Причем имманентная взаимосвязь культуры и информации в значительной степени опосредована их связью с эволюционными процессами, где на всём протяжении супермагистрали глобальной эволюции именно информация играет главенствующую роль, а культура – на её определённом этапе. Энергия играет здесь «подчинённую» роль, она необходима для эволюции, но не является столь приоритетной и определяющей, как информация и соответствующие информационные процессы.

«Информационная стрела» глобальной эволюции, вырвавшись за пределы биологического мира, увлекла за собой остальные свойства и характеристики материи, без которых информация в принципе существовать не может, т.е. энергию, вещество, пространство, время и т.д. Культура вначале не была выделена на ранних этапах антропосоциогенеза из формирующейся синкретической человеческой деятельности, в том числе, из синкретизма триады – вещество, энергия и информация. Она появилась тогда, когда возникла эволюционная необходимость развития этого относительно автономного преимущественно информационного социального процесса для дальнейшего продолжения глобальной эволюции.

В отечественной литературе информационный подход к культуре в советское время развивал эстонский академик Ю.М. Лотман, который дал определение понятия человеческой культуры как «совокупности всей ненаследственной информации, способов ее организации и хранения» [14, с. 146]. Возможность создания, накопления и передачи негенетическим путем разнообразной информации другим индивидам и потомкам принципиально отличает человека от его диких родственников. В дальнейшем удалось обобщить эти идеи и обосновать обсуждаемое здесь положение – что социальная ступень эволюции характеризуется особой надиндивидуальной системой средств накопления, хранения, переработки и передачи информации от поколения к поколению, что важно для коллективного объединения входящих в общество индивидов.

Информационная трактовка культуры уже вошла и в учебные пособия. А.С. Кармин дает «теоретическое определение культуры», исходя из того, что «культура – это социальная информация, которая сохраняется и накапливается в обществе с помощью создаваемых людьми знаковых средств» [15, с. 17]. При этом он подчеркивает, что культура представляет собой особый тип информационного процесса, которого не знает природа, поскольку, например, у животных информация сохраняется и накапливается в самом организме, а ее передача от одного поколения к другому происходит, в основном, генетическим путем.

В связи с информационной трактовкой культуры, возникает ряд вопросов, которые имеют отношение к различным концепциям природы информации. Очевидно, что, например, семиотическое представление культуры выделяет лишь знаковые средства, с помощью которых кодируется социальная информация, расшифровываются тексты и другие носители информации. Вместе с тем, ряд ученых не считают, что культура состоит только из знаков, которые только обозначают другие объекты. В этом случае культура может представляться в качестве информационного феномена, но не в семиотическом понимании информации, а в более широкой – атрибутивной трактовке.

Тем самым культура представляет собой лишь те информационные процессы и системы, которые, будучи артефактами либо естественными объектами и процессами, оказались вместе с тем и знаками, которые человек наделил смыслом и ценностью. Культура выступает в форме социально-информационных, и прежде всего информационно-семиотических процессов, в отличие от информационно-генетических процессов в мире живой природы.

Информация, заключенная в культуре, имеет двойственный характер: в артефактах часть информации не зависит от человека и человечества, а другая часть, благодаря интеллектуально-духовной дея-

тельности человека, наделена смыслом. Кроме того, сознание человека может наделять смыслом не только артефакты, но и естественные феномены, что сплошь и рядом имеет место в естественных науках.

Таким образом, культура представляет собой лишь те информационные процессы и системы, которые, будучи артефактами либо естественными объектами и процессами, оказались вместе с тем символами и знаками, наделенными человеком смыслом, который уже может быть отделен от ценности. Культура тем самым выступает в качестве социально-информационных и, прежде всего, информационно-семиотических процессов, в отличие от информационно-генетических процессов в живой природе. Но если это принять, то приоритетно-первичным в появлении смыслов (значений) выступает не сам этот смысл, а его создание сознанием человека, поскольку культура не существует без человека. И это уместно подчеркнуть, поскольку имеется точка зрения, что «не наличие у людей сознания отличает их от своих «родственников» – крупных травоядных животных, хотя оно гораздо более развитое и включает механизм мышления, а возможность создания, накопления и передачи внегенетическим путем разнообразной информации другим индивидам и потомкам» [16, с. 30-31].

Наличие надбиологических механизмов, т.е. программ, кодов, алгоритмов и т.д., играет важнейшую роль в развитии общества, выражая не только его отличие от биологической ступени, но и фактически глубинную информационную сущность социальной ступени развития, которую уместнее было бы назвать социально-культурной ступенью. Многие важные тенденции социального развития можно объяснить, исходя из того, что природа социального заключена именно в культуре.

Это свойство внутренне связано с сознанием и одухотворенными им различными видами человеческой деятельности и не только орудийной. В обществе одно без другого не существует, причем именно сознание наделяет смыслом артефакты-знаки, которые в дальнейшем активно участвуют в формировании сознания, как индивидов, так и коллективного интеллекта в форме культуры. Сознание человека в становлении и развитии культуры играет роль опережающего фактора, без которого не состоялся бы социально-информационный процесс. Принцип экзогенного накопления культурной информации, имеющий информационно-синергетический характер, вовсе не отменяет положения о том, что эта информация обретает свою культурную форму лишь в ходе придания сознанием человека смысла артефактам и иным объектам, после чего они включаются в семиосферу как главную область человеческой культуры.

Таким образом, можно констатировать наличие социокультурного принципа экзогенного накопления, хранения, передачи и преобразования информации. И это согласуется с принципами синергетики, согласно которым рост информации в эволюционирующей материальной системе происходит за счет окружающей среды, за счет изъятия у нее негэнтропии и иных ресурсов. Но при этом происходит процесс «окультуривания» этой среды, в которой повышается либо понижается информационное содержание сферы социально-культурной деятельности. Поэтому важно понять, что замещение пространства с «естественной информацией» на пространство с «искусственной информацией» должно происходить преимущественно в местах с низким уровнем «естественной информации», в основном – в неживой природе. Но это стало очевидным только в наше время, когда глобально-экологический кризис стал угрожать человечеству близкой катастрофой.

Вероятнее всего, в основном, благодаря действию экзогенного информационно-культурного принципа, биологическое развитие человека, отражаемое в геноме, приостанавливается или существенно замедляется, становится в каком-то смысле второстепенным для дальнейшей эволюции социальной ступени. Основная информационная деятельность, т.е. накопление и движение информации, уже вынесена за пределы человеческого организма, тем самым не так сильно стала влиять на сам этот организм. Во всяком случае, за последние 40 тыс. лет геном человека изменился меньше, чем на 0,02%, и человечество как биологический вид практически уже не изменяется в различных природных условиях планеты (при этом генетическое отличие шимпанзе как наиболее близкого примата к человеку составляет около 1%) [17]. Это в определённой степени произошло потому, что, формируясь в основном под действием цефализационных процессов, человек передал основную функцию накопления информации в социосфере от своего организма внешнему для него, но сущностно связанному с ним культурогенезу.

Феномен социального (как культурного) появляется тогда, когда происходит «вынос» основных информационных процессов (накопления, хранения, преобразования информации) за пределы структурного элемента социальной ступени. Это специфическая и сущностная характеристика социальной ступени эволюции, связанная с культурогенезом, выражает ее принципиальное отличие от биологической ступени. Поэтому вряд ли можно полностью согласиться с тезисом, что: «Культура – это особое проявление природы человека, специфическая форма реализации его генетически унаследованной социальности...» [18, с. 37]. Несомненно, что генетически унаследованная социальность играла и до сих пор играет определённую роль в становлении и реализации феномена культуры. Но эта роль абсолютизируется, когда культура видится только как система социального поведения человека и групповой коммуникации, являющейся развитием соответствующих программ поведения животных, и когда заранее делается акцент на исследовании общего между животными и человечеством.

В уже состоявшемся культурогенезе, а тем более в современной культуре, всё-таки главную роль играют не эндогенные генетические процессы накопления и движения информации, а экзогенные генетические свойства и характеристики освоения информации, которые начали развиваться благодаря цефализации. Именно эти информационные свойства и механизмы культурогенеза как процесса позволили преодолеть геоцентризм предшествующей социальной

эволюции и выйти за пределы планеты, чего в принципе не смогли бы сделать любые животные с их «привязанностью» к эндогенно-генетическим процессам.

Культурогенез, тем самым, обретает не только свою информационную сущность, но и поистине космическое значение, распространяясь за пределы планеты. А это уже связано с перспективами выживания человечества не только на Земле, но и в пространствах Вселенной, на что ориентировал в свое время основоположник теоретической космонавтики К.Э. Циолковский, а в настоящее время к этому космическому расселению активно призывают американский предприниматель Илон Маск и всемирно известный английский физик Стивен Хокинг.

В отличие от предыдущих ступеней эволюции материи человечество начинает расширять сферу своего распространения сначала на Земле, а затем и в Космосе, не только (и не столько) для получения вещественно-энергетических ресурсов, но и, прежде всего, для продолжения своих информационных процессов, и накопления информации в расширяющейся социосфере. Феномен расширения социальной ступени уже именуется Большим социальным взрывом, и он имеет глубинную культурно-информационную природу. Культурно-цивилизационный процесс распространяется не только по пространству планеты, но в дальнейшем – и внеземному пространству, где происходит дальнейшее всё более масштабное и ускоренное овладение информацией и негэнтропией окружающей природной среды. По сути, это уже как социокультурный, так и социоприродный эволюционный процесс, который означает, что дальнейшая самоорганизация материи будет осуществляться через культуру, составляющую «ядро» человеческой цивилизации.

ЗАКЛЮЧЕНИЕ

Рост информационного содержания материальных систем в процессах непрерывной и длительной самоорганизации имеет определенное и необратимое направление, своего рода «информационную стрелу» эволюции, которая особым образом проявляется в цефализационных процессах и культурогенезе. Эта векторность глобальной эволюции характерна лишь для информационных процессов самоорганизующихся материальных систем, но не для масс-энергетических, пространственных и иных известных нам характеристик эволюционных процессов. Именно при переходе от биологической к социальной эволюции существенно возрастает роль информации и информационных процессов, которые, выделяясь из синкретического триединства вещества, энергии и информации (но пока не доминируя над ними), начинают играть всё большую роль в продолжении супермагистрали глобальной эволюции.

Проблемы информации и информационных процессов в культуре и культурогенезе пока не занимают должного места в культурологических и информационных исследованиях. Несколько лет назад был поставлен вопрос о формировании информационной культурологии, которая в широком понимании мыслится как интерпретация и исследование феномена культуры на основе понятия информации и информационного подхода [19, 20]. Между тем, на наш взгляд, становление информационного подхода в культурологии и развитие информационной культурологии как научного направления позволят создать единую теоретико-методологическую базу изучения культуры и существенно усилят ее прогностическую функцию.

Но поскольку культура и культурогенез теперь рассматриваются и в информационном ракурсе, то важно заглянуть в их будущее. Эта цель ставилась и раньше, когда речь шла о становлении и перспективах информационного общества, которое может рассматриваться как одна из ступеней ноосферогенеза (инфоноосфера) в том случае, если одновременно будет осуществляться переход на «устойчивую» траекторию дальнейшей эволюции.

СПИСОК ЛИТЕРАТУРЫ

1. Bazaluk O.A. The theory of evolution // Philosophy and Cosmology. – 2015. – Vol. 1. – P. 25-33.
2. Базалук О.А. Теория эволюции: От космического вакуума до нейронных ансамблей и в будущее: Монография. – Киев: МФКО, 2014. – 312 с.
3. Урсул А.Д. Освоение космоса (Философско-методологические и социологические проблемы). – М.: Мысль, 1967. – 276 с.
4. Урсул А.Д., Урсул Т.А. Универсальный эволюционизм (концепции, подходы, принципы, перспективы). – М.: РАГС, 2007. – 326 с.
5. Ильин И.В., Урсул А.Д., Урсул Т.А. Глобальный эволюционизм: идеи, проблемы, гипотезы. – М.: Московский университет, 2012. – 616 с.
6. Вернадский В.И. Научная мысль как планетное явление / отв. ред. А. Л. Яншин. – М.: Наука, 1991. – 271 с.
7. Тейяр де Шарден П. Феномен человека. – М: АСТ, 2012. – 381 с.
8. Буровский А.М. Феномен мозга. Тайны 100 миллиардов нейронов. – М.: Яуза; Эксмо, 2010. – 247 с.
9. Красилов В.А. Нерешенные проблемы теории эволюции. – Владивосток, 1986. – 91 с.
10. Кассирер Э. Философия символических форм: В 3 т. / пер. с нем. С.А. Ромашко. – М.-СПб.: Университетская книга, 2002.
11. Основы государственной культурной политики. – М.: Минкультуры, 2015. – 72 с. – URL: http://mkrf.ru/upload/mkrf/mkdocs2016/OSNOVI-PRINT.NEW. indd.pdf (дата обращения: 20.06.2017).
12. Стратегия государственной культурной политики на период до 2030 года. – М., 2016. – 45 с. – URL: http://government.ru/media/files/AsA9RAyYVAJnoBuKgH0qEJA9IxP7f2xm.pdf (дата обращения: 20.06.2017).
13. Уайт Л. Избранное: эволюция культуры: пер. с англ. – М.: Российская политическая энциклопедия (РОССПЭН), 2004. – 1064 с.
14. Лотман Ю.М. Статьи по семиотике культуры и искусства / предисл. С.М. Даниэля; сост. Р.Г. Гри-

горьева. – СПб.: Академический проект, 2002. – 543 с.
15. Кармин А.С. Культурология: учеб. пос. – СПб: Питер, 2009. – 240 с.
16. Лосев К.С. Мифы и заблуждения в экологии. – М.: Научный мир, 2010. – 223 с.
17. Боринская С.А. О генетических отличиях человека и шимпанзе // Антропогенез.ру. – URL: http://antropogenez.ru/article/75/ (дата обращения: 20.06.2017).
18. Флиер А.Я. Происхождение культуры: новая концепция культурогенеза // Знание. Понимание. Умение. – 2012. – № 4. – URL: http://www.zpu-journal.ru/e-zpu/2012/4/Flier_The-Origin-of-Culture/ (дата обращения: 20.06.2017).
19. Колин К.К., Урсул А.Д. Информационная культурология. Предмет и задачи нового научного направления. – Саарбрюккен: Lambert academic publishing, 2011. – 249 с.
20. Колин К.К., Урсул А.Д. Культура и информация. Введение в информационную культурологию. – М.: Стратегические приоритеты, 2015. – 300 с.

Материал поступил в редакцию 20.06.17.

Сведения об авторе

УРСУЛ Аркадий Дмитриевич – доктор философских наук, профессор, заслуженный деятель науки РФ, академик Академии наук Молдавии, почётный работник высшего профессионального образования РФ, директор Центра глобальных исследований и профессор факультета глобальных процессов Московского государственного университета им. М.В. Ломоносова, Москва.
e-mail: ursul-ad@mail.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 001.103 – 028.27

А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев

Интенсивное использование цифровых данных в современном естествознании*

Анализируются общие подходы и технологии, связанные с хранением и обработкой цифровых данных в различных дисциплинах. Показано, что, вне зависимости от специфики предметной области, работа с обширными массивами данных, полученных в результате эксперимента или моделирования, требует одинакового методического обеспечения, включая процедуры курирования, поддержку метаданных, аннотирование данных сведениями об их генезисе и качестве. Как пример дисциплины интенсивно использующей цифровые данные, рассмотрена междисциплинарная область «свойства веществ и материалов». Проанализированы новые подходы к задачам интеграции разнородных данных по свойствам, способные учесть вариации в структуре данных в зависимости от выбора класса веществ, состояния образца, условий эксперимента и других факторов.

Ключевые слова: *наука о данных, курирование данных, качество данных, анализ данных, свойства веществ и материалов, материаловедение, теплофизика*

ВВЕДЕНИЕ

Понятие об особой форме науки, максимально ориентированной на работу с цифровыми данными, возникает примерно в 90-е гг. XX в. с введением в оборот ряда примерно равнозначных терминов: *data-intensive* (*data-dominated, data-centric*) science, *eScience* и т.п. Заметим, что эта лексика распространилась заметно раньше, чем более общее понятие о «Больших данных», имеющее несколько иной смысл [1]. Во-первых, речь идет именно о науке как производителе и потребителе данных, в то время как проблема больших данных в основном затрагивает сферу бизнеса и общественной жизни. Во-вторых, для науки, ориентированной на данные, ключевой характеристикой является не столько объем данных, сколько весь характер деятельности. Пожалуй, наиболее яркий признак науки о данных – это практическое доминирование работы с ними, масштаб которой заметно превышает масштабы эксперимента, вычислений и моделирования. Появился даже специальный термин *четвертая парадигма* [2, 3], под которой имеется в виду именно работа с данными, в то время как предыдущие три охватывали эксперимент, теорию и моделирование (рис. 1).

При всем многообразии научных дисциплин с характерными для них типами данных можно выделить общие факторы, инициировавшие сдвиг в практике и технологии научного исследования:

1) резкий рост возможностей научных инструментов (телескопов, ускорителей, спутников и т.п.), обеспечивших беспрецедентное производство первичных данных;

2) появление высокопроизводительных средств и технологий, автоматизирующих процессы сбора, обработки и распространения данных;

3) непрерывный рост публикационного потока, поставляющего вторичные, т.е. прошедшие экспертизу и обработку данные в табличном, графическом или аналитическом виде.

Результаты измерений, поступающие от спутников, телескопов, ускорителей и прочих научных инструментов, собранные для хранения и анализа в масштабных базах данных обладают огромной ценностью для исследователя, однако создают немало проблем при организации рабочего процесса. Существенно, что это связано не столько с необходимостью разработки и внедрения новых программно-аппаратных средств, сколько с необходимостью перестройки всего научного подхода, причем вне зависимости от конкретной дисциплины.

В проекте глобальной инфраструктуры исследовательских данных GRDI2020[1] [4] указывается, что наука с доминированием данных неизбежно влечет новый, ориентированный на данные (data-centric) подход к концептуализации, организации и проведению всего исследовательского процесса.

* Работа выполнена при поддержке Российского научного фонда, грант № 14-50-00124.

[1] GRDI2020 (Towards a 10-Year Vision for Global Research Data Infrastructures) – 10-летний международный проект экосистемы глобальных исследовательских данных, реализуемых в рамках ЕС (см. http://www.grdi2020.eu).

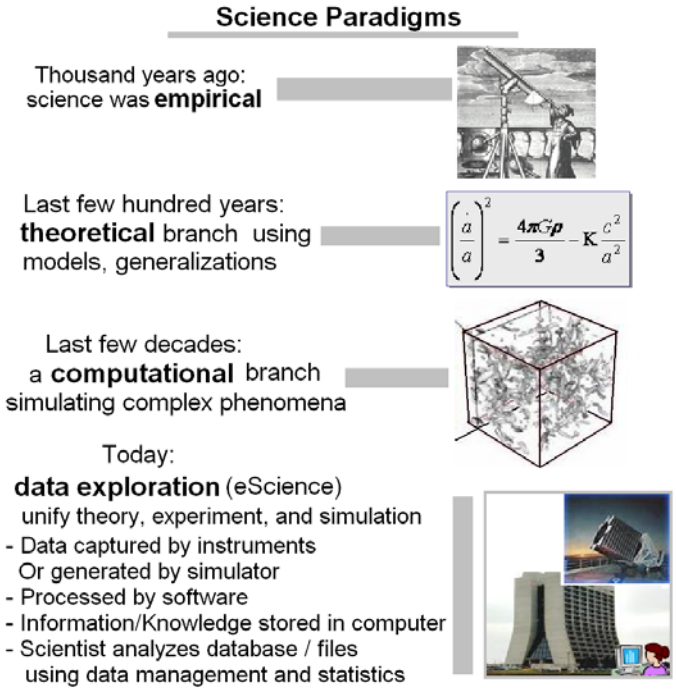

Рис. 1. Схематичное представление роли и места четвертой научной парадигмы [3]

Наиболее заметно это проявилось в науках о Земле и космосе (климатология, астрономия и др.), медико-биологических дисциплинах и материаловедении. Оформились целые направления с однотипным названием *X-информатика*, где X – выделяет конкретную область знания, со своими типами данных и задачами. Среди наиболее развитых можно указать гео- и биоинформатику [3, 5], но появились и другие, с более узкой тематикой, как например астро- [6] или наноинформатика [7]. В то же время в научном сообществе сложилось представление о внутреннем единстве этих дисциплин, свидетельством чего стало множество публикаций, конференций и т.д., посвященных общему предмету под условным названием eScience. Поскольку, однако, архивирование данных с их систематизацией и обработкой всегда занимали в науке важное место, нужно понять границы или масштаб конкретной дисциплины, начиная с которых ее оправданно относить к дисциплинам, ориентированным на данные.

Междисциплинарный характер работе с данными был придан, когда в 1966 г. образовался CODATA (Комитет по численным данным для науки и техники), а с 2002 г. –специальный журнал «Data Science Journal» (http://datascience.codata.org/), охватывающий множество дисциплин и методических вопросов: технологии публикаций, стандартов обмена данными и т.п. На его страницах впервые была сформулирована концепция о единой науке данных как самостоятельной академической дисциплине [8]. Однако на протяжении ряда десятилетий проблематика «данных» оставалась на периферии общего внимания. Поэтому активный переход, начавшийся примерно с 2000 г., к новой единой методологии, условно названной «четвертой парадигмой», требует объяснения, какие сдвиги в предмете и практике исследований повлекли очередную научную революцию. Возникает и второй вопрос: на каком основании под общим понятием объединяют столь разные предметы, как био-, гео- или астроинформатика, в чем проявляются единство подходов и их реализация?

ПРИЗНАКИ И ГРАНИЦЫ НАУКИ С ИНТЕНСИВНЫМ ИСПОЛЬЗОВАНИЕМ ДАННЫХ

В литературе можно найти множество дефиниций, упоминающих атрибуты eScience, как то: масштабное хранение и распространение данных, сетевая инфраструктура, охват всех стадий исследования от постановки задачи до анализа результатов. Например, Wikipedia приводит фрагменты дефиниции, данной компьютерной журналисткой S. Bohle [9], которая в качестве признаков eScience называет применение компьютерных технологий, включая подготовку, накопление и распространение данных в сочетании с их долгосрочным хранением и доступом ко всем материалам исследования. В аналогичном определении на сайте конференции IEEE (https://escience-conference.org/) в качестве характерных признаков выделены охват всех стадий исследования от постановки задачи до анализа результатов, а также наличие специальной инфраструктуры. Автор работы [10] подчеркнул важность *data-centric* стиля документирования научного процесса с повсеместным переходом от ссылок на публикации к ссылкам на наборы данных. Наиболее подробно проблема рассмотрена в лекции Джима Грея (jimgray.azurewebsites.net/), (одного из ведущих экспертов eScience), где он сформулировал концепцию «четвертой парадигмы», декларировав, что ведущей и универсальной тенденцией современной науки становится использование компьютера как средства хранения и обработки данных,

вне зависимости от их формата, предметного содержания и генезиса (см. рис. 1).

Авторы этих формулировок, однако, избегают четкой дефиниции с выделением формальных признаков, которые могут объяснить, когда та или иная дисциплина может рассматриваться как eScience. Ближе всего к точному и адекватному определению науки о данных подошли авторы из КНР [11]. В качестве ее первого признака они признали **объект исследования**, подчеркнув, что объектом выступают именно данные как элемент киберпространства, а не предметы окружающего мира. Практически это означает, что для eScience работа с данными поглощает больше ресурсов (финансовых, кадровых, временных), чем те, которые тратятся на эксперимент или моделирование. Это определяет и доминирующую роль данных по сравнению с такими артефактами исследования как образец или оборудование.

Следуя логике авторов [11], в качестве второго признака, дополняющего объект исследования, оправданно принять **метод исследования** с опорой на базу данных (БД) как основной инструмент, наряду с обширной инфраструктурой в виде онтологий, контролируемых словарей, средств Semantic Web и т.п.

Наконец, третьим непременным признаком eScience является наличие мощной «аналитики», т. е. программной инфраструктуры, способной обеспечить обработку данных, их статистический анализ, выявление закономерностей и т.п. Универсальность, проявляемая в близости/сходстве ассортимента методик, слабо зависящего от предметной области, позволяет рассматривать такую инфраструктуру как один из типовых признаков науки с интенсивным использованием данных.

Аналитика работы с данными включает две группы методов. Первая группа под собирательным названием *data mining* [12, 13] (интеллектуальный анализ данных) объединяет совокупность методов обнаружения в больших объемах данных ранее неизвестных трендов и закономерностей. Их основу составляют различные подходы к классификации, моделированию и прогнозированию. Среди них деревья решений, нейронные сети, генетические алгоритмы, методы нечеткой логики и др. К этой же группе принято сейчас относить и многочисленные статистические методы, например, регрессионный и факторный анализ, анализ временных рядов, непараметрическую статистику и др. Вторая группа методов дополняет возможности data mining средствами визуального анализа, который всегда был неотъемлемой частью исследования. Условная схема на рис. 2 показывает типовую стратегию анализа с последовательным сокращением объема данных – от просмотра исходных данных к многомерному анализу и наконец, к методам data mining.

Применительно к eScience, с характерным для нее ростом объема и усложнением структуры данных, возникают новые потребности в задачах визуализации. Среди них – визуализация потока данных, генерируемых в ходе крупномасштабного моделирования природных или технологических процессов; визуальный интеллектуальный анализ с выдачей рекомендаций; визуализация с отслеживанием информации об эволюции процесса и источниках данных на предыдущих этапах и ряд других [14]. Используемые для анализа (как визуального, так и ориентированного на выявление закономерностей) программно-аппаратные средства имеют широкую сферу применения, что само по себе усиливает междисциплинарный характер науки о данных.

Перечисленные три признака, а именно использование **данных как основного объекта**, **БД как основного инструмента** и наличие мощной встроенной **аналитики**, достаточны для четкого выделения eScience как принципиально нового типа организации научного исследования. Интересно, что при детализации различных этапов работы с данными, четко выявляется междисциплинарный характер eScience, позволяя использовать сходные по целям, хотя и различные по содержанию подходы.

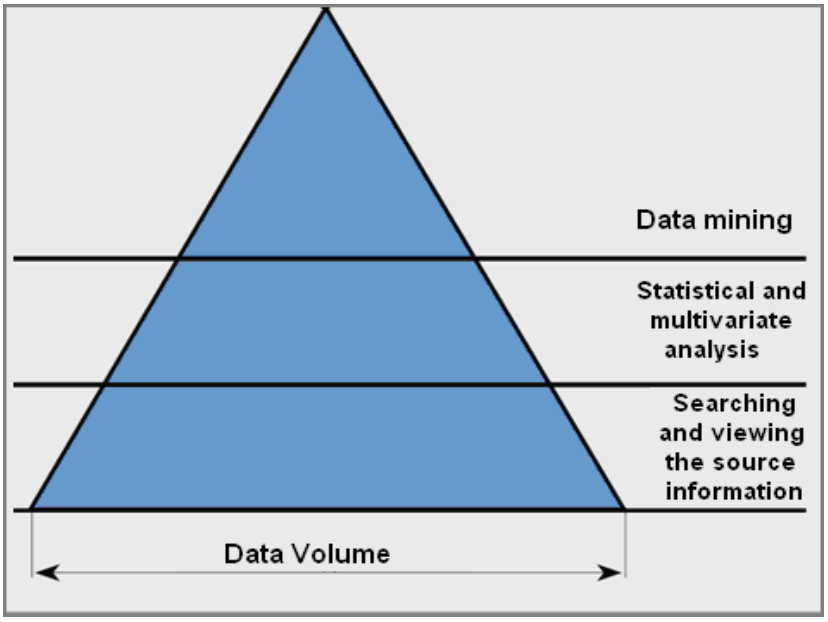

Рис. 2. Типовая стратегия анализа данных [12, 13]

КУРИРОВАНИЕ ДАННЫХ

Цифровые данные, как и объекты архивного или библиотечного хранения, никогда не лежат «мертвым грузом», требуя перманентных процедур, гарантирующих сохранность и пригодность к использованию. Их совокупность составляет предмет особой деятельности под названием **курирование данных** (*data curation*). В отечественной литературе этот термин почти не встречается, уступая по распространенности более привычному, «систематизация», что конечно имеет иной смысл, а именно: упорядочивание данных в соответствии с определенной рубрикацией. За рубежом понятие курирования цифровых данных возникло тоже не слишком давно, примерно с 90-х годов. Его происхождение связано с музейной практикой, где издавна использовалась работа куратора, включающая заботу о сохранении образцов, их обновлении, описании и т.п. [15]. В явном виде термин *data curation* прозвучал в статье Diane Zorich's [16], автор которой пришла к выводу о единстве проблем, присущих библиотекам, музеям, и научным центрам, занятым формированием и поддержкой цифровых архивов. Согласно ее концепции, наборы цифровых данных, как и поддерживающие их средства (словари, тезаурусы, метаданные) должны регулярно контролироваться и обновляться для согласованности данных, сохранения их качества, доступности и т.п., а деятельность в этом направлении составляет суть *курирования*.

Проблемы курирования данных оказались настолько сложны, что потребовалось создание журнала International Journal of Digital Curation (www.ijdc.net/) и профильного Центра (Digital Curation Centre, www.dcc.ac.uk/), издающих методические материалы в помощь экспертам, занятым сбором и хранением данных. Центром разработана специальная модель (The DCC Curation Lifecycle Model), формализующая состав и последовательность отдельных элементов курирования. Для динамических наборов данных курирование включает непрерывное расширение или обновление за счет включения ссылок на аннотации, сетевые ресурсы и электронные публикации. Среди основных целей курирования данных, как правило, называют их сохранение, описание, меры безопасности, так называемую очистку, т.е. контроль и восстановление качества, и ряд других.

Хранение данных (preservation). В значительной степени долгосрочное хранение проблема сугубо технологическая, связанная с ограниченным сроком жизни устройств памяти (не более ста лет), причем это требует достаточно заметных ресурсов. На сегодняшний день сохранение данных включает перенос архива каждые несколько лет на новые носители и конверсию устаревающих форматов в более современные. Предложено несколько стратегий долговременного хранения: миграция, эмуляция и использование, так называемого, виртуального компьютера. Миграция предполагает последовательную смену форматов по мере изменения технологий. Эмуляция, предложенная Д. Ротенбергом (www.clir.org/pubs/reports/rothenberg) позволяет воссоздать устаревшее аппаратное и программное обеспечение оригинального устройства на современной платформе, что является идеальным (хотя и дорогостоящим) способом электронного архивирования. Наконец, концепция виртуальной машины, специально спроектированной для хранения архивов, комбинирует методы миграции и эмуляции без аппаратного обеспечения. Машина конвертирует исходную форму данных в универсальный технологически независимый формат, наподобие XML.

Помимо сложных инженерных решений, помощь в сохранности данных может оказать соблюдение ряда условий хранения. Одно из них – использование качественных метаданных и стандартных форматов, типа текстового или XML при безусловном отказе от специфических, имеющих хождение в узком кругу пользователей. Другое условие - исключение избыточных данных, поскольку сокращение их объема естественно упрощает и снижает расходы на хранение. Если избыток данных порожден использованием определенных моделей, то на длительный период достаточно хранить исходные для моделирования данные и сведения о процедуре их обработки. Например, обширные таблицы теплофизических данных можно изъять из хранилища, сохранив лишь исходные для расчета данные, например параметры уравнения состояния [13]. Другой пример, иллюстрирующий потенциал возможной экономии на объеме хранения, – молекулярно-динамические эксперименты, итогом которых являются обширные протоколы, отражающие эволюцию координат и импульсов для большого ансамбля частиц. Сохранив сведения о параметрах потенциала, а также все программные инструменты для моделирования, можно рассчитывать на воспроизводимость результатов тех же численных экспериментов в неопределенном будущем, сняв нагрузку, связанную с хранением обширных данных моделирования.

Описание данных. Помимо физической сохранности, важнейшим из условий долгосрочного использования, является качественное описание. Сами по себе цифровые данные, вне зависимости от их происхождения и структуры, никогда не бывают самоописываемы. Метаданные документируют контекст, включая сведения о том, как данные были получены, когда и кем введены для хранения, каким процедурам обработки и ревизии подвергались за период хранения. Имеется обширная литература по научным метаданным и специфике их использования в различных дисциплинах [17, 18]. Сопровождая предметную информацию, метаданные позволяют: идентифицировать набор данных с указанием его положения в хранилище; определить правила доступа; описать логическую структуру и форматы данных; обеспечить работу всевозможных средств анализа и визуализации.

Созданы стандарты метаданных для разных дисциплин и типов документа, собранные в директории (rd-alliance.github.io/metadata-directory/). Там же приведены стандарты, пригодные для любой из дисциплин, например **Dublin Core metadata set –** для

формального описания цифровых ресурсов или **DataCite Metadata Store** – для их идентификации и цитирования.

В наших работах [17, 19, 20] детально описаны метаданные для теплофизических свойств (**ThermoML**), характеристик обычных материалов (**MatML**) и объектов нанотехнологий.

Происхождение данных (*provenance*) – одна из наиболее важных групп метаданных, включающих сведения о создании, лицензировании и версионности. Они всегда генерируются при создании новой версии и указывают соотношения между двумя версиями набора данных. В частности, метаданные могут включать название программы, создавшей новую версию, и идентификаторы других наборов данных или ресурсов, использованных при ее создании. Информация о происхождении собирается на протяжении всего жизненного цикла в процессе курирования. Представление и обмен информацией о происхождении данных позволяет: отслеживать эволюцию данных; выявлять наличие ошибок или устаревших данных, а также ответственных за это программных агентов; исправлять обнаруженные ошибки, распространяя поправки на весь жизненный цикл. Для ее формализации создан ряд стандартов, например **PROV** (www.w3.org/TR/prov-overview/).

Курирование научных данных всегда предполагает какие-либо формы их аттестации, т.е. оценки достоверности, научной ценности, возможности использования и т.п. Оценки могут исходить как от создателей наборов, так и внешних экспертов, отвечающих за хранение или использующих данные при анализе. Общий принцип аттестации примерно таков же, как и в работе с публикациями, где оценки расставляются как автором, так и в процессах *peer review*. Специфика обширных наборов данных проявляется в формализации этого процесса и принятии некоторых междисциплинарных стандартов. Результаты аттестации проявляются при совместной обработке различных наборов на этапе их анализа, а также при определении необходимых сроков хранения.

Оценка неопределенности. Первое и обязательное требование к научным данным – предоставление информации об их неопределенности, как для исходных («сырых») данных, так и полученных в ходе последующей обработки [17]. Для разъяснения общего понятия о неопределенности привлекаются близкие, но не идентичные термины: неточность, неполнота, погрешность, несогласованность, двусмысленность. Неопределенность в сырых данных вносится за счет случайных или систематических ошибок, присущих прибору и методу измерения. Дополнительным источником неопределенности становятся метод обработки данных, средства прогнозирования и оценки, недостоверность модели и т.п. Характеристики, используемые при оценке неопределенности, достаточно многообразны. Простейший способ – «удвоить» данные, приписывая к значению свойства X в каждой точке значение среднеквадратичной погрешности σ. Именно этот способ предложил Стоунбрейкер, один из ведущих специалистов в теории БД, ссылаясь на общепринятое использование нормального распределения результата измерений [21]. Во многих случаях, однако, оправдано принять более сложную и детализированную характеристику неопределенности. Например, в стандарте метаданных **ThermoML** [19, 22], аттестующих исходные и рассчитанные данные по теплофизическим свойствам, выделено три типа оценок: стандартные σ, расширенные σ_L и комбинированные. Первые две могут быть приписаны зависимым и независимым переменным, последний тип объединяет оба вклада.

Вариант оценки неопределенности связан также с полнотой «первичных» данных и требованиями к качеству данных со стороны приложений. Во многих случаях достаточно ограничиться одной оценкой на весь набор (в абсолютных или относительных величинах), в других значения неопределенности принимаются различными для отдельных величин из набора данных или даже для отдельных точек. Так, статистические оценки в стандарте **ThrmoML** [22] приписаны каждой точке из набора данных, а для набора в целом предложена группа ориентировочных оценок достоверности: повторяемость, воспроизводимость, отклонение от аппроксимирующей кривой. Выбор типа неопределенности требует введения соответствующего элемента в набор научных метаданных.

Оценка качества данных. Качество данных – более сложное понятие, чем неопределенность, не имеющее четкого и однозначного определения. Традиционно считается, что качество данных объединяет такие характеристики как точность, полнота и согласованность. Однако, обеспечивая потребности пользователей, данные должны удовлетворять типовым требованиям качества продукта: доступность, удобство в использовании, своевременность в предоставлении и т.п. Практически в каждой из дисциплин предлагались собственные критерии, основанные на специфических требованиях к их достоверности, полноте, форме представления и т.п. Так, например, предложенная в работе [23] процедура аттестации данных о наноматериалах принимала во внимание полноту описания образцов, методов измерений или оценок, обоснованность определения неопределенности и ряд других факторов, каждый из которых оценивался индикатором качества. Большой опыт архивирования, поддержки и анализа научных данных лег в основу единой системы оценки качества, где из множества характеристик, присущих разным дисциплинам, удалось выявить универсальный набор из 15 критериев, распределенных по четырем базовым категориям: точность, релевантность, представление и доступность. Три последние категории в большей степени характеризуют не столько сами данные, сколько степень их соответствия запросам пользователя и удобству его работы. По выражению авторов [24], совокупность этих категорий и относящихся к ним критериев, подчинена ключевому условию “fitness for use”.

Авторы работы [25], с учетом специфики больших данных, расширили этот набор, предложив иерархическую систему показателей (табл. 1). Система

включает пять базовых категорий, каждая из которых представлена несколькими элементами. Интегральной мерой достоверности является категория *надежность*, включающая, наряду с точностью, такие элементы как согласованность, целостность, полнота и проверяемость. Одним из аспектов надежности или достоверности можно рассматривать *правдоподобие*, раскрываемое рядом условий (см. табл. 2). Остальные категории и элементы отражают «товарные» характеристики, включая условия доступа, качество сервиса, спецификацию, т.е. документирование ресурса и качество метаданных.

Таблица 1

Категории и элементы в системе оценки качества данных [25]

Доступность	**Удобство в использовании**	**Надежность**	**Релевантность**	**Качество представления**
Доступ к данным	Документирование	Точность	Соответствие требованиям пользователя	Читабельность
Своевременность	Правдоподобие	Целостность		Наличие структуры
Авторизация[1]	Наличие метаданных	Согласованность		
		Полнота		
		Проверяемость		

[1] Наличие прав или лицензий на использование данных

Таблица 2

Индикаторы некоторых элементов в системе оценки качества данных [25]

Элементы	**Индикаторы**
Доступ к данным	Наличие интерфейса доступа к данным
	Легкость доступа к открытым данным и оплаты коммерческого доступа
Правдоподобие	Источником данных являются авторитетные организации
	Регулярная проверка корректности данных экспертами
	Нахождение данных в диапазоне известных или общепринятых значений.
Точность	Обеспечена точность самих данных
	Представление данных отражает истинное состояние источника информации
	Представление данных не допускает многозначной трактовки
Целостность	Форматы данных очевидны для пользователя и соответствуют стандартам
	Данные согласованы с требованием структурной целостности
	Данные согласованы с требованием целостности контента
Согласованность	Концепции, области определения и форматы данных не меняются в ходе обработки
	Данные остаются согласованными и проверяемыми в течение определенного интервала времени
	Обеспечена согласованность и проверяемость с данными из других источников
Соответствие требованиям пользователя	Собранные данные, хотя не полностью соответствуют теме, освещают ее определенный аспект
	Тематика большинства запрошенных наборов данных соответствует тематике запроса

Чтобы раскрыть и детализировать смысл каждого из элементов в табл. 1, они сводятся к нескольким индикаторам, под которыми понимаются подтверждаемые или неподтверждаемые суждения. В табл. 2 для иллюстрации приведены индикаторы, раскрывающие некоторые из элементов качества. Например, точность данных включает не только обычную информацию о погрешности, но и наличие данных об источнике и требование ясности представления, исключающее неоднозначность в трактовке. Тем самым, невозможность подтвердить второе и третье утверждение должна рассматриваться как фактор, снижающий точность, вне зависимости от заданной погрешности. Аналогично, по нескольким критериям раскрываются и другие элементы, например целостность данных, где предполагается их соответствие модели данных и невозможность модификаций в ходе жизненного цикла.

Совокупность из 14 элементов, представленных в табл. 1, совместно с их индикаторами, позволили авторам [25] разработать модель многоаспектной оценки качества данных. Проведенная аттестация (по данной или аналогичным методикам) ставит перед кураторами данных задачу их контроля на соответствие критериям качества. Этот процесс, получивший название очистка данных (*cleaning* или *cleansing*), включает выявление и исправление искаженных, неточных или неполных данных. Загрязнение данных возникают по разным причинам, среди которых: ошибки ввода и дублирования, старение записей, неправильное распределение данных по полям. Искаженные данные резко снижают их качество, делая по существу непригодными для анализа, что передает принятая в литературе поговорка «мусор на входе, мусор на выходе».

Очистка данных не допускает снижения их качества в ходе хранения за счет контроля за выполнением ряда критериев, таких как точность, согласованность, полнота и др. (см. табл. 1). Естественно, что формальные процедуры не могут сказаться на таких критериях, как правдоподобие или полнота метаданных, выполнение которых зависит от всего контекста: наличие новых данных, изменение требований к результатам исследования и т.п. В процессе очистки проводят идентификацию, замену, модификацию или удаление некорректных или неверно отформатированных записей. В отдельных случаях возможно лишь повторное дублирование материала. При этом стараются создать одну новую версию, даже если в дальнейшем она должна храниться в нескольких разрозненных системах.

Помимо указанной процедуры очистки (*cleaning, cleansing*), есть альтернативная процедура (*data purging*) примерно с той же задачей, а именно удаление старых или бесполезных элементов данных. Отличие состоит лишь в том, что в этом случае целью удаления является высвобождение места для новых данных, в то время как исходная процедура (*data cleaning*) фокусируется на повышении точности данных с удалением синтаксических или типографских ошибок.

СВОЙСТВА ВЕЩЕСТВ И МАТЕРИАЛОВ

Особенности предметной области. Наряду с науками о Земле или медициной, интенсивное использование данных все более захватывает междисциплинарную сферу исследований под названием **свойства веществ и материалов**. Сбор и систематизация данных по свойствам вещества издавна занимают важное место в физике, химии, материаловедении и других дисциплинах. Хороший пример – издаваемые с XIX в. знаменитые справочники Бельштейна по свойствам органических и Гмелина по неорганическим веществам [26]. Вне зависимости от конкретной области, эта деятельность включает примерно одинаковые этапы: (1) компиляция «сырых» данных; (2) экспертиза с выявлением их согласованности, неопределенности и т.п. (3) статистическая обработка с определением параметров регрессионных зависимостей; (4) распространение рекомендуемых данных в виде публикаций и/или заполнение баз данных (БД).

Тенденция к доминированию информатики при обобщении свойств вещества отмечалась уже в нашей работе [13]: «…по объемам накопленных и публикуемых данных, охвату библиографии и т. д., справочно-аналитическая деятельность все более приобретает черты… бизнес-процесса, требующего не только надежной научной базы, но и отработанной технологии контроля и управления информационными потоками. В этих условиях потребность в информационных технологиях возникает на всех стадиях справочной деятельности, включая накопление и экспертизу данных, их распространение, координацию совместной работы и проч.».

В отличие от таких областей как астрономия или медицина, здесь источником данных является нарастающий публикационный поток, отражающий синтез новых веществ, их лабораторное изучение и моделирование. При этом, объем данных определяется не столько числом изучаемых объектов, сколько безграничным многообразием условий синтеза, измерений, морфологических и микроструктурных особенностей. В формировании потока больших данных из трех признаков (так называемых, «3V – Volume, Velocity, Variety» [3]) именно последний, т. е. разнообразие типов и источников данных, играет решающую роль.

Примерно с 60-х годов прошлого века основным итогом в процессе подготовки данных о свойствах стало создание БД, сильно различающихся по назначению, объему и функциональным возможностям. С увеличением их числа и объема хранимых данных в научном сообществе возник консенсус относительно необходимости интеграции данных и стандартизации процессов обмена. Различные подходы к решению этой проблемы в отношении теплофизических свойств были рассмотрены в ряде работ [19, 27].

В то же время, в отличие от наук о Земле и медицины, в междисциплинарной области **свойства веществ и материалов** еще не сложились единые подходы к сбору и хранению данных, что позволило бы ее квалифицировать как eScience. Авторы обзора [28] выделили целую группу тормозящих факторов,

которые относятся к специфике работы научного сообщества. Среди них: слабое восприятие лексики, привычной для специалистов по информационным технологиям; различные предпочтения в выборе методов поиска и оценки новых данных; несовпадение целей хранения и поиска данных для разных областей, к примеру, индустриальной химии или технологии конструкционных материалов; недостаточная распространенность стандартов и структурированных данных; недостаток финансовой поддержки проектов интеграции и стандартизации данных. Можно сказать, что традиционно специалисты по свойствам предпочитают работу с *малыми данными*, т. е. автономными ресурсами (БД, электронными справочниками и др.), что резко противоречит сложившейся культуре работы с интегрированными данными, которая доминирует в науках о Земле или медицине. Поэтому, при обсуждении концепций и технологий работы с данными по свойствам вещества, многие авторы настойчиво рекомендуют заимствовать успешный опыт, достигнутый в области био- или геоинформатики [28, 29].

Вещества и материалы – общность и различия данных. При общем сходстве работы с данными по свойствам, специфика конкретной дисциплины с неизбежностью проявляется в выборе форматов, моделей данных и наконец, требований к единой инфраструктуре, соединяющей хранилища, сервисы и аналитические средства. Весь мир объектов, для которых собираются и обрабатываются данные по свойствам, несколько упрощая, можно разделить на вещества, материалы и, наконец, наноструктуры. Чистое вещество определяется стехиометрической формулой и фазовым состоянием, раствор или смесь - химическим и фазовым составом и дисперсностью.

Принципиальное отличие материала от вещества состоит в том, что стехиометрия и/или состав недостаточны для его идентификации. Приходится принимать во внимание множество факторов, связанных с особенностями технологии, состоянием образца, методами измерения и тестирования, данными об изготовителе и заводской марке и т.п. Все эти факторы с еще большей силой проявляются во взаимодействии с наноматериалами, причем за счет технологий молекулярного уровня, приходится учитывать также многообразие топологий и морфологических признаков наноразмерных объектов. Естественно, что физические различия указанных категорий сказываются и на типологии данных, прежде всего, за счет их объема и сложности структуры, которые нарастают по мере перехода от чистых веществ к традиционным материалам (стали, сплавы, керамика и т.п.) и далее к наноматериалам. При этом структура данных усложняется как за счет привлечения множества экстрафакторов (технология, состояние образца и т.п.) в дополнение к параметрам состояния, так и за счет вариации номенклатуры свойств при изменении класса материала. По-видимому, впервые проблема сложной структуры данных о свойствах, обсуждалась в нашей работе [19].

Применительно к чистым веществам и растворам, на решение этой проблемы, т. е. формализацию и унификацию представления данных сложной структуры, был направлен международный проект IUPAC **ThermoML** [27]. Разработанная процедура предусматривала объединение численных данных и метаданных в рамках одного документа, написанного посредством языка XML (язык расширяемой разметки), что представляет новую форму публикации, промежуточную между неструктурированным текстом и структурированными таблицами в БД. Ее своеобразие в том, что она одинаково доступна для «прочтения» как человеком, так и компьютером. С рядом физико-химических журналов («Journal of Chemical and Engineering Data», «Journal of Chemical Thermodynamics», «Fluid Phase Equilibria») авторы проекта достигли договоренности о выделении из публикаций комплекса численных данных для их распространения в формате **ThermoML**. Отвлекаясь от деталей, изложенных в специальной литературе (см. ссылки в статье https://ru.wikipedia.org/wiki/XML), заметим, что XML обеспечивает универсальный формат для представления структурированных документов различного класса, в качестве которых могут рассматриваться, например, числовые таблицы, математические или химические тексты. Стандарт **ThermoML** обеспечил разные возможности представления теплофизических данных: охват свыше 120 свойств для индивидуальных веществ, растворов и химических реакций; поддержку иерархических структур; ключевую роль метаданных; сведения об особенностях образца, методах измерений и очистки, способах оценки и т.п.

Однако, несмотря на декларированную универсальность, схема **ThermoML** не приобрела такого распространения, чтобы можно было говорить о появлении единой инфраструктуры. По-видимому, сказались и упомянутые выше барьеры [28], побуждающие специалистов по свойствам предпочесть работу с «малыми данными» в ущерб переходу к глобальной инфраструктуре, интегрирующей данные разного формата и структуры. Как следствие, при переходе от обычной теплофизики к материаловедению, возникла потребность в разработке новых средств и технологий, способных к интеграции данных значительно большего объема и с большим разнообразием структуры.

Информатика в области материалов (materials informatics). Значимость материалов как основы производства, быта и прочих сфер человеческой жизни определяет и важную роль соответствующей X-информатики, в рамках которой могут возникнуть общие подходы к систематизации научных данных. Тенденция к переходу от автономных ресурсов к интегрированным средам в материаловедении возникла достаточно давно. В отечественной литературе эти проблемы рассмотрены в работах [30, 31].

Еще в 1985 г. было объявлено о создании в США Национальной сети данных, которая включала целый набор БД по свойствам разных материалов [32]. Для обмена и распространения данных был предложен единый стандарт **MatML** (www.matml.org/history.htm), напоминающий его аналог **ThermoML**, но учитывающий, что объектами являются не вещества с известной стехиометрией, а материалы, свойства которых существенно зависят от технологии изготов-

ления, факторов внешнего воздействия и т.п. Как следствие, язык не имеет жестко фиксированной схемы, пригодной на все случаи, как в **ThermoML**, допуская за пользователем неограниченные возможности создавать собственные теги. Элементы **MatML** определяют небольшой набор типовых понятий, таких как <Property-Data>, <Name>, <Units> и т.п. Включены также элементы, определяющие происхождение (источник) данных, например, <Metadata> или <DataSourceDetails>. **MatML** был спроектирован так, чтобы удовлетворить любые потребности исследований или разработок без жесткой привязки к определенному типу объектов или приложений, Соответственно **MatML**-документы характеризуются высокой степенью вариабельности по отношению к сфере приложений.

В качестве альтернативной стратегии объединения независимых БД по материалам использовалось онтологическое моделирование, позволяющее включить многообразные типы данных с явным определением их семантики (смысла). Подробный анализ возможностей этого направления дан в работах [33-35]. В работе [35] показано, как, основываясь на онтологии свойств, можно спроектировать БД, а в дальнейшем перестраивать ее структуру, ориентируясь на вновь открытые объекты и их характеристики. На базе онтологий, как доступных в сети, так и вновь разработанных, технологии Semantic Web позволяют построить и распространить в сети документы с данными для конкретного материала, которые попадают в так называемое пространство связанных данных. Этот способ сетевого распространения данных обеспечивает их автоматическую интеграцию с родственными документами, созданными из других ресурсов, но относящихся к тем же материалам и/или свойствам. В рамках CODATA еще в 2006 г. была создана специальная группа под названием “Exchangeable Materials Data Representation to support Scientific Research and Education” [34], задачей которой было распространение методов онтологического моделирования для интеграции материаловедческих данных.

Качественный скачок в развитии материаловедческой информатики произошел с принятием в США новой программы под названием Material Genome initiative (MGI, mgi.nist.gov/). Ее основная цель – существенно обновить весь цикл работ по созданию новых материалов, с двукратным ускорением их разработки и внедрения в промышленность. На смену длительным и дорогим процедурам эмпирического поиска, MGI предусматривает активное привлечение методов компьютерного моделирования, использующих как физические принципы, так и аналитику больших данных, способную выявить закономерности типа «структура-свойство» или «структура-обработка-свойство».

Соответственно на одно из ведущих направлений, наряду с экспериментом и моделированием (рис. 3), выдвинулась технология цифровых данных, ориентированная на задачи их хранения, обработки и распространения. Ключевым моментом, отличающим новый этап в развитии *materials informatics,* является создание, взамен отдельных ресурсов, масштабной инфраструктуры, объединяющей под единым управлением репозитарии данных, сервисы доступа к ним и средства анализа и визуализации данных.

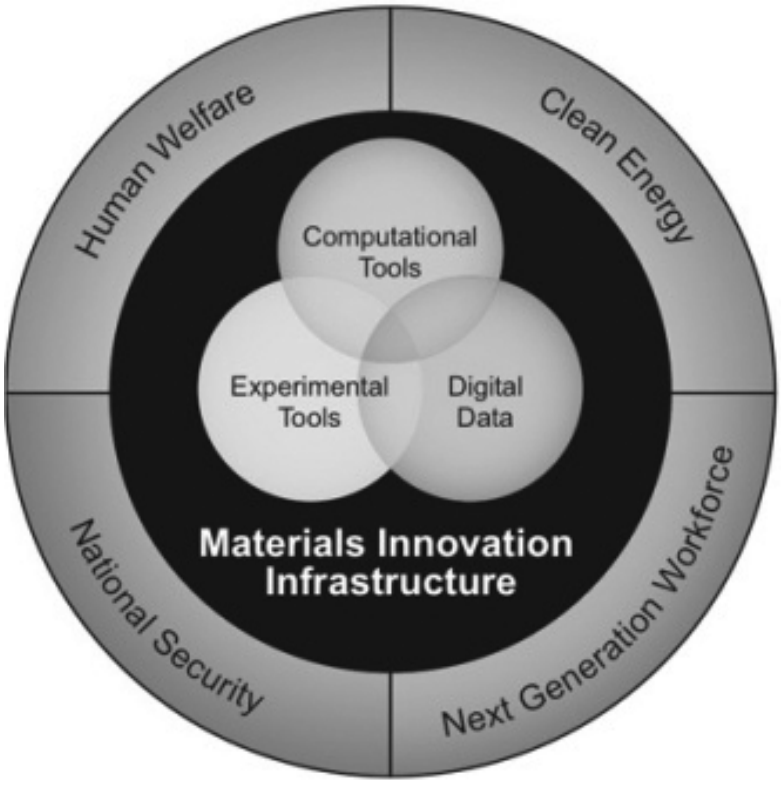

Рис. 3. Основные элементы программы MGI – эксперимент, вычислительные средства и цифровые данные (см. mgi.nist.gov/2014-mgi-strategic-plan-0)

Целью подобных инфраструктур, согласно программе MGI, является агрегирование, хранение и распространение данных различного формата, как структурированных (из реляционных БД), так и неструктурированных, например текстов или изображений. Конкретная задача, решаемая в ходе создания каждой инфраструктуры, включает стандартизацию форматов и типов данных, метаданных, критериев для ввода и архивирования данных, протоколов, необходимых для свободного обмена данными. Другой элемент инфраструктуры – «федеративные БД», т. е. распределенная сеть репозитариев с достаточно развитым набором метаданных для идентификации данных, оценки их качества и поддержки семантических запросов в пределах всей сети БД. Интересно, что, согласно стратегическому плану MGI (mgi.nist.gov/2014-mgi-strategic-plan-0), прообразом создания инфраструктуры данных по материалам должны стать разработки, ранее нашедшие применение в проекте Генома человека (www.ornl.gov/hgmis) и в геоинформационной системе EarthCube (www.nsf.gov/geo/earthcube).

К настоящему времени созданы или находятся в процессе разработки несколько подобных инфраструктур [28, 36, 37]. Ведущая роль в разработке концепций и технологий принадлежит NIST (Национальный институт стандартов, США), организации, имеющей давний опыт в создании и распространении критически оцененных данных, относящихся к свойствам вещества. Прототип инфраструктуры, предлагаемой NIST [36], включает две базовые составляющие: систему курирования данных (Materials Data Curation System, MDCS) и регистр ресурсов (NIST Materials Resource Registry, NMRR). Первая из них (система курирования) представляет собой платформу для сбора данных и конверсии их к некоторому единому стандарту. Вторя составляющая – регистр ресурсов, децентрализованная сеть, обеспечивающая поддержку и распространение ресурсов произвольного типа, от публикаций до коллекций структурированных данных. Подробное описание всех компонентов инфраструктуры можно найти на сайте (mgi.nist.gov/) , где собраны документы, относящиеся к работам NIST в рамках программы MGI.

Система курирования данных (mgi.nist.gov/materials-data-curation-system)**.** В инфраструктуре NIST курирование обеспечивает работу с данными о материалах, исходно представленными во многих форматах. Система позволяет вести загрузку, структурирование, хранение и распространение данных различного происхождения: экспериментальных, расчетных, из публикаций и т.д. Интеграцию разнотипных данных из разных источников удается преодолеть за счет подбора соответствующих метаданных и конверсии к новому формату, пригодному для массового использования. В качестве такового использован XML, ранее успешно опробованный в проекте теплофизических данных **ThermoML** [27]. Его применение позволяет структурировать данные и метаданные в рамках текстового файла, доступного для восприятия как человеком, так и компьютером. Пользователю предоставлен набор шаблонов со схемой XML, т. е. детальным описанием структуры метаданных, пригодных для определенного типа данных. Первый шаг в курировании – выбор из заготовленного набора конкретного шаблона в качестве контейнера для загрузки данных (экспериментальных или результатов моделирования) вместе с метаданными, рис. 4.

Как только шаблон выбран, пользователь вводит данные в форму, основанную на отобранной XML схеме. Предлагаемые пользователю шаблоны (XML-схемы) охватывают наиболее типичные варианты ввода данных. Чтобы выйти за эти границы, предлагается специальный «составитель шаблонов» (*Template Composer*) для формирования непредусмотренной структуры. При этом наличие метаданных обеспечивает возможность эффективного поиска с хорошо определенным смыслом.

Каждый из файлов, сам по себе представляющий автономную информационную единицу, загружается в БД, хорошо приспособленную к работе с большими и сложно структурированными данными. В качестве таковой выбрана популярная БД MongoDB (www.mongodb.com/what-is-mongodb), относящаяся к категории БД, специально ориентированных на работу в эру больших данных. Наличие БД позволяет преодолеть естественное ограничение XML, как текстового файла, обеспечивая независимое хранение связанных с текстом изображений и других нетекстовых объектов, например PDF или медиафайлов.

Одно из главных преимуществ XML – заложенная в нем способность к трансформации в другие форматы, например CSV (comma-separated values) для числовых таблиц, хорошо приспособленные для загрузки в табличный процессор Microsoft Excel или реляционную БД. После завершения курирования, т.е. создания скорректированных XML-файлов, охватывающих исходно полуструктурированные данные, система пригодна к эксплуатации: подачи запросов, их объединения, поиска и сохранения нужных результатов.

Система курирования уже успешно использовалась в ряде проектов [29], среди которых БД NanoMine, ориентированная на свойства и технологии нанокомпозитов, репозитарий межатомных потенциалов, систему сбора данных, полученных в ходе компьютерного моделирования материалов и другие средства.

Регистр ресурсов (NIST Materials Resource Registry, NMRR). При наличии множества ресурсов по свойствам материалов (репозитариев, БД и т.п.) доступ к ним достаточно затруднен из-за различия принятых схем, терминологии, типов данных и т.д. Задача регистра **NMRR (**mgi.nist.gov/materials-resource-registry) – преодолеть существующее многообразие, обеспечив прямую и эффективную связь пользователя с ресурсом посредством системы метаданных. Введенные в регистре метаданные связаны со спецификой содержания, включая класс материала, методы сбора данных, сведения о микроструктуре материала и т.п. Сами метаданные распространяются по сети, образуя единый каталог. Тем самым, регистр обеспечивает функционирование децентрализованной сети, с поддержкой и распространением ресурсов произвольного типа, от публикаций до коллекций структурированных данных. Обе составляющие цифровой инфраструктуры NIST взаимно дополнительны, поскольку система курирования используется, чтобы

сделать данные по материалам доступными, а регистр включает их в общую децентрализованную среду, обеспечивая доступ к ним и совместное использование, наряду со всей совокупностью накопленных материаловедческих данных.

Пользователь регистра, желая зарегистрировать свой ресурс (архив, БД и др.), может использовать шаблоны метаданных, принятые в узком сообществе, например, среди специалистов по керамикам или полимерам. Кроме того, есть система автоматического сбора метаданных из существующих ресурсов, со специальным протоколом, известным как OAI-PMH (www.openarchives.org/pmh/). Протокол сбора метаданных (OAI-PMH, version 2.0) создан организаций открытых архивов (Open Archives Initiative) с целью их интеграции и свободного обмена данными (первоначально – для свободного обмена препринтами публикаций). Протокол обеспечивает механизм, посредством которого администраторы ресурсов раскрывают их метаданные за счет их отображения к универсальному формату Dublin Core (http://dublincore.org/). В итоге, сбор метаданных, проводимый согласно протоколу, позволяет работать, совершая поиск и обмен данными среди множества архивов. Формируя набор данных посредством системы курирования, пользователь может внести его в регистр, обеспечив возможности децентрализованного поиска.

Формат полуструктурированных данных. Одной из общих особенностей, присущих данным по физическим свойствам, является сложная и подверженная вариациям логическая структура данных, зависящая от выбора класса веществ и множества непредсказуемых факторов, связанных с технологией, состоянием образца, внешней средой и т.п. Работа с подобными данными, которые принято называть *полуструктурированными*, составляет одну из основных задач, возлагаемую на инфраструктуру, объединяющую данные для широкого круга объектов. Ее эффективное решение было найдено при создании политематической платформы Citrination (citrination.com), где собрана информация из множества источников (экспериментальных, расчетных, сведений из публикаций и технических документов) для материалов всех классов. Это одна из наибольших в мире коллекций, включающая не менее 3 млн записей с открытыми данными. При этом обеспечена возможность работы с двумя типами данных, как структурированными (записями из БД), так и неструктурированными (графики, PDF-файлы и т.п.).

Наибольший интерес представляет использованный в проекте новый формат, наилучшим образом приспособленный к данным сложной структуры, что позволяет решить давнюю проблему подстройки хранилища данных к вариациям их структуры.

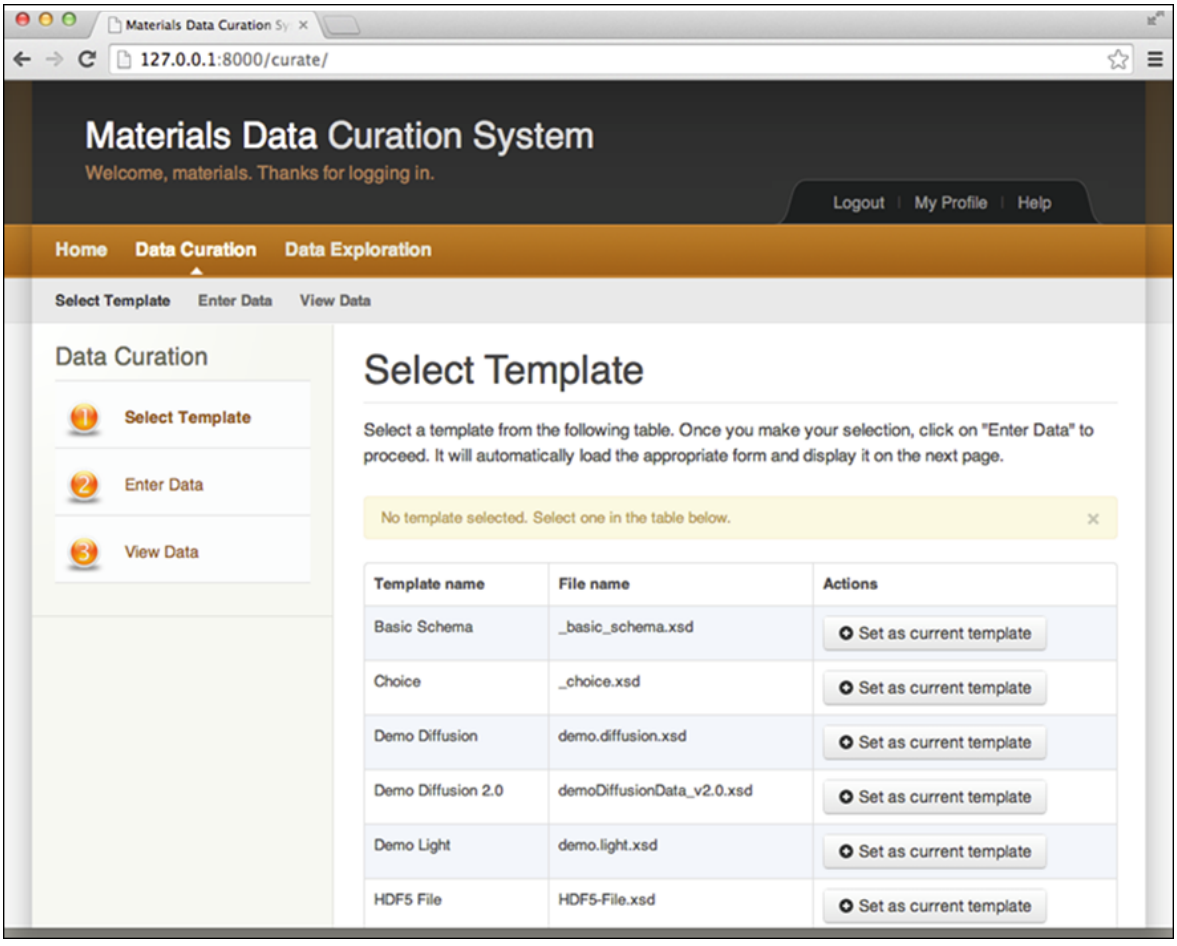

Рис. 4. Курирование данных – интерфейс с отбором шаблонов со схемой данных [36]

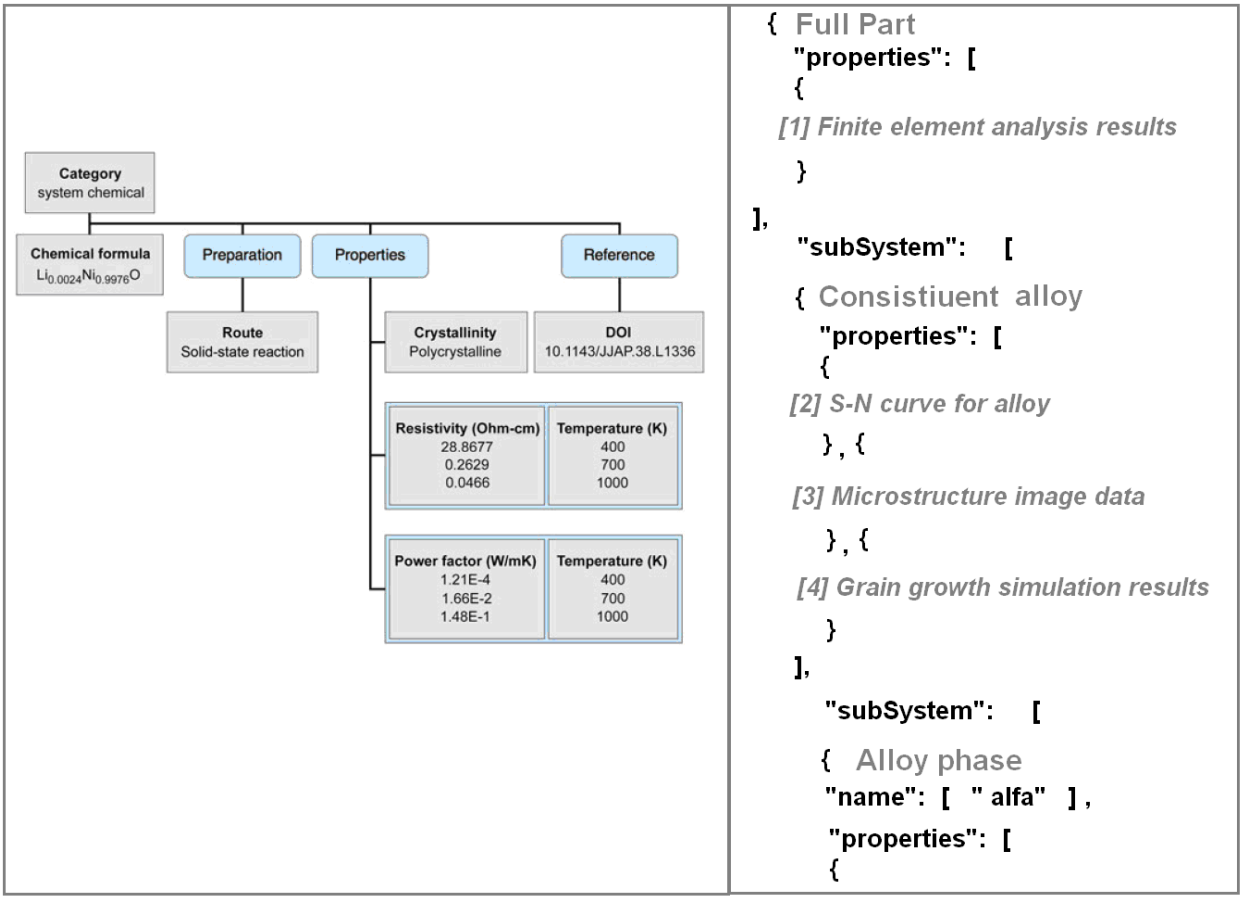

Рис. 5. Примеры данных сложной структуры в системе Citrination:
слева – блок-схема данных по свойствам термоэлектрика при заданной химической формуле и методе производства; справа – PIF-схема, включающая ссылки на графические файлы с результатами расчетных и экспериментальных исследований хрупкого алюминия [37]

Специально предложенный для платформы формат, под названием **Physical information file** (PIF) [37], так же, как и XML, является текстовым форматом обмена, доступным для чтения человеком, но имеющим ряд преимуществ при охвате широкого диапазона веществ, их свойств и методов производства. Общий принцип, выдержанный при проектировании PIF-схемы, состоял в том, чтобы по возможности обеспечить три условия – *общность, гибкость и легкость структурирования данных*, облегчая стадии загрузки данных за счет усложнения стадии их анализа.

Одно из условий общности схемы – наличие обязательных элементов, определяющих название, метод производства и свойства любого из объектов, включенных в схему. Общность схемы выдерживается также за счет ее иерархической структуры, что позволяет детально раскрыть содержание каждого из элементов, например название или метод производства. Другое качество PIF-схемы, гибкость, достигается возможным включением дополнительных полей и объектов, которые отсутствовали в априорно заготовленной схеме, что позволяет ее подстраивать к особенностям предметной области.

Приведенные на рис. 5 примеры из работы [37] иллюстрируют богатый потенциал формата, способного передать как свойства обычного вещества, определяемого химической формулой, так и сложную графическую информацию о хрупкости алюминия, включая изображения микроструктуры и кристаллической структуры, вычисленной квантово-механическим методом.

ВЫВОДЫ

Рассмотрена одна из общих тенденций современной науки – максимальная ориентация на работу с цифровыми данными, поступающими в результате мониторинга природной среды, лабораторного эксперимента или компьютерного моделирования. Как следствие, практика и технология современного исследования приобрели сходные черты для разных дисциплин: наук о Земле и астрономии, биологии и медицины, теплофизики и материаловедения. Общность подходов связана с необходимостью поддержки и активной работы с обширными массивами данных, включая их хранение, организацию доступа, использование аналитических средств обработки и визуализации. Масштаб работы обусловил появление

новых форм и методов работы с данными, таких как курирование, назначение и стандартизация метаданных, регламентация архивного хранения, оценивание неопределенности и качества данных. В статье подробно описаны эти формы со ссылками на различные ресурсы (публикации, web-сайты), где приведены методические материалы и стандарты.

Детально изучена междисциплинарная область «свойства веществ и материалов», которая может также рассматриваться как eScience из-за объема и сложности обработки данных. Основным источником информации здесь является публикационный поток, отражающий перманентное появление и изучение новых объектов: веществ, материалов, наноструктур. Объем и структура данных по свойствам определяются многообразием как самих объектов, так и технологий их синтеза, методов измерений, влияний внешней среды и других факторов. Их совокупность оказывается наиболее значимой в сфере материаловедения, что породило, так называемую, *material informatics*, в рамках которой создаются масштабные инфраструктуры, в задачу которых входит интеграция и распространение разнородных данных. При их создании удалось приблизиться к решению давней проблемы поддержки полуструктурированных данных, отражающих вариации номенклатуры и типа свойств в зависимости от конкретного класса материалов.

СПИСОК ЛИТЕРАТУРЫ

1. Lynch C. Big data: How do your data grow? // Nature. – 2008. – Vol. 455. – P. 28-29.
2. Gray J., Szalay A.S., Thakar A.R. et al. Online Scientific Data Curation, Publication, and Archiving // Technical Report MSR-TR-2002-74. Microsoft Research.
3. The Fourth Paradigm. Data-Intensive Scientific Discovery / eds. T. Hey, St. Tansley, Kr. Tolle. 2009, Microsoft Corporation.
4. Thanos C. A Vision for global research data infrastructures // Data Science Journal. – 2013. – Vol. 12. – P. 71-90.
5. Zhao J., Corcho O., Missier P. et al. eScience // Handbook of Semantic Web Technologies. – Berlin Heidelberg: Springer-Verlag, 2011. – P. 703-733.
6. Borne K. Astroinformatics: Data-Oriented Astronomy Research and Education // Earth Science Informatics. – 2010. – Vol. 3, № 1. – P. 5-17.
7. Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Трахтенгерц М.С. Наноинформатика: задачи, методы и технологии // Научно-техническая информатика. Сер. 1. – 2016. – № 10. – С. 1-18; Erkimbaev A.O., Zitserman V.Yu., Kobzev G.A., Trakhtengerts M.S. Nanoinformatics: Problems, Methods, and Technologies // Scientific and Technical Information Processing. – 2016. – Vol. 43, № 4. – P. 199-216.
8. Smith F.J. Data Science as an academic discipline // Data Science Journal. – 2006. – Vol. 5. – P. 163-164.
9. Bohle S. What is E-science and How Should it be Managed // SciLogs, 12 June 2013.
10. Erbach G. Data-centric view in E-science inforation systems // Data Science Journal. – 2006. – Vol. 5. – P. 219-222.
11. Zhu Y., Xiong Y. Towards Data Science // Data Science Journal. – 2015. – Vol. 14, № 8. – P. 1–7.
12. Забежайло М.И. Интеллектуальный анализ данных – новое направление развития информационных технологий // Научно-техническая информация. Сер. 2. – 1998. – №5. – С. 6-17.
13. Зицерман В.Ю., Кобзев Г.А., Фокин Л.Р. Перспективы развития информационно-аналитических средств в задачах сбора и генерации справочных данных // Научно-техническая информация. Сер. 1. – 2004. – № 2. – С. 7-14.
14. Hansen C., Johnson C. R., Pascucci V., Silva C. T. Visualization for data-intensive science // In: "The Fourth Paradigm. Data-Intensive Scientific Discovery" / eds. T. Hey, St. Tansley, Kr. Tolle. – Microsoft Corporation, 2009. – 153 p.
15. Palmer C.L., Weber N.M., Muñoz T., Renear A.H. Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data // Archive Journal. – 2013. – Iss. 3.
16. Zorich D.M. Data management: Managing electronic information: Data curation in museums // Museum Management and Curatorship. – 1995. – Vol. 14, № 4. – P. 431.
17. Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А. Роль метаданных в создании и использовании информационных ресурсов о свойствах веществ и материалов // Научно-техническая информация. Сер. 1. – 2008. – № 11. – С. 13-19.
18. Хохлов Ю.Е., Арнаутов С.А. Обзор форматов метаданных. Портал «Российские электронные библиотеки». – URL: www.elbib.ru/index.phtml?page=elbib /rus/methodology/md_rev
19. Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Фокин Л.Р. Логическая структура физико-химических данных. Проблемы стандартизации и обмена численными данными // Журнал физической химии. – 2008. – Т. 82, № 1. – С. 20-31.
20. Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Трахтенгерц М.С. Универсальная система метаданных для характеристики наноматериалов // Научно-техническая информация. Сер. 1. – 2015. – № 10. – С. 8-20; Erkimbaev A.O., Zitserman V.Yu., Kobzev G.A., Trakhtengerts M.S. A Universal Metadata System for the Characterization of Nanomaterials // Scientific and Technical Information Processing. – 2015. – Vol. 42, № 4. – P. 211-222.
21. Stonebraker M. et al. Requirements for science data bases and sciDB. In: "Fourth Bienial Conference on Innovation Data Systems Research", CIDR 2009. – URL: www-db.cs.wisc.edu/cidr/cidr2009/Paper_26.pdf
22. Chirico R.D., Frenkel M., Diky V.V. et al. ThermoMLs: An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property

Data. 2. Uncertainties // J. Chem. Eng. Data. – 2003. – Vol. 48, №5. – P. 1344-1359.
23. Елецкий А.В., Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Трахтенгерц М.С. Теплофизические свойства наноразмерных объектов: систематизация и оценка достоверности данных // Теплофизика высоких температур. – 2012. – Т. 50, № 4. – С. 524-532.
24. Wang R.Y., Strong D.M. Beyond Accuracy: What Data Quality Means To Data Consumers//Journal of Management Information Systems. – 1996. – Vol. 12, № 4. – P. 5-33.
25. Cai L., Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era // Data Science Journal. – 2015. – Vol. 14, № 2. – P. 1-10.
26. Потапов В.М., Кочетова Э.К. Химическая информация. Где и как искать химику нужные сведения. – М.: Химия, 1988. – 224 с.
27. Frenkel M. Global communications and expert systems in thermodynamics: Connecting property measurement and chemical process design // Pure Applied Chem. – 2005. – Vol. 77, № 8. – P. 1349-1367
28. Hill J., Mulholland G., Persson K.et al. Materials science with large-scale data and informatics: Unlocking new opportunities // MRS Bulletin. – 2016. – Vol. 41, № 5. – P. 399-409.
29. Hunt W.H., Jr. Materials Informatics: Growing from the Bio World // JOM. – 2006. – Vol. 58, №7. – P. 88.
30. Киселева Н.Н., Дударев В.А. Инфраструктура обеспечения данными специалистов в неорганической химии и материаловедении // Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016. – С. 191-198.
31. Дударев В.А. Интеграция информационных систем в области неорганической химии и материаловедения. – М.: КРАСАНД, 2016. – 320 с.
32. Rodgers John R., Cebon D. Materials Informatics // MRS Bulletin. – 2006. – Vol. 31, № 12. – P. 975-980.
33. Еркимбаев А.О., Жижченко А.Б., Зицерман В.Ю., Кобзев Г.А., Сон Э.Е., Сотников А.Н. Интеграция баз данных по свойствам вещества. Подходы и технологии // Научно-техническая информация. Сер. 2. – 2012. – №8. – С. 1-8; Erkimbaev A.O., Zhizhchenko A.B., Zitserman V.Yu., Kobzev G.A., Son E.E., Sotnikov A.N. Integration of Databases on Substance Properties: Approaches and Technologies // Automatic Documentation and Mathematical Linguistics. – 2012. – Vol. 46, № 4. – P. 170-176.
34. Zhang X., Zhao C., Wang X. A survey on knowledge representation in materials science and engineering: An ontological perspective // Computers in Industry. – 2015. – Vol. 73. – P. 8–22.
35. Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Косинов А.В. Связывание онтологий с базами данных по свойствам веществ и материалов // Научно-техническая информация. Сер. 2. – 2015. – №12. – С. 1-16.
36. Dima A., Bhaskarla S., Becker C. et al. Informatics Infrastructure for the Materials Genome Initiative // JOM. – 2016. – Vol. 68, № 8. – P. 2053-2064.
37. Michel K., Meredig B. Beyond bulk single crystals: A data format for all materials structure–property–processing relationships // MRS Bulletin. – 2016. – Vol. 41, № 8. – P. 617-623.

Материал поступил в редакцию 18.04.17.

Сведения об авторах

ЕРКИМБАЕВ Адильбек Омирбекович – кандидат технических наук, зав. лабораторией теплофизических баз данных Объединенного Института высоких температур РАН (ОИВТ РАН), Москва
e-mail: adilbek@ihed.ras.ru

ЗИЦЕРМАН Владимир Юрьевич – кандидат физико-математических наук, ведущий научный сотрудник лаборатории теплофизических баз данных ОИВТ РАН
e-mail: vz1941@mail.ru

КОБЗЕВ Георгий Анатольевич – доктор физико-математических наук, главный научный сотрудник, советник НИЦ электрофизики и тепловых процессов ОИВТ РАН
e-mail: gkbz@mail.ru

УДК 004.627 : 519.217

Е.В. Ларкин, А.В. Богомолов, А.Н. Привалов

Методика оценивания временн*ы*х интервалов между транзакциями в алгоритмах сжатия речевых сообщений*

Изложена методика оценивания временных интервалов между транзакциями в алгоритмах сжатия речевых сообщений на основе сложного марковского процесса, каждое состояние которого представляет собой 2-параллельный марковский процесс, описывающий «соревнование» источника сигнала, заполняющего буфер, и приемника сигнала, опорожняющего буфер. Сложный марковский процесс преобразован в ординарный процесс, состояния которого моделируют количество заполненных в текущий момент ячеек буфера, что позволило получить зависимость, связывающую вероятность отказа, объем буферной памяти и математические ожидания времен заполнения и опорожнения буфера.

Ключевые слова: *сжатие речевых сообщений, алгоритм сжатия, обработка аудиоданных, марковский процесс, буферизация данных*

ВВЕДЕНИЕ

Программное сжатие речевых сообщений широко применяется в системах передачи данных [1, 2]. В алгоритмах сжатия речевых сообщений программный модуль сжатия размещается между сенсорным блоком, решающим задачу восприятия звуковых колебаний и их преобразования в цифровой код, и аппаратурой передачи данных, решающей задачу трансляции сжатого звукового сигнала по узкополосному каналу передачи данных. В силу жестких ограничений на стабильность частоты транзакций по обоим указанным интерфейсам, и особенностей обработки сигналов с помощью контроллеров на базе ЭВМ фон-неймановской архитектуры, передаваемые данные буферизируются. Необходимость оптимизации использования объема буферной памяти определяет актуальность разработки методики оценивания временных интервалов между транзакциями в алгоритмах сжатия речевых сообщений, что являлось целью проведенного исследования.

ОБЩАЯ МОДЕЛЬ ФУНКЦИОНИРОВАНИЯ СИСТЕМЫ ПЕРЕДАЧИ АУДИОДАННЫХ СО СЖАТИЕМ

Типовая система передачи аудиоданных со сжатием приведена на рис. 1а. В неё входят: источник сигнала, контроллер программной обработки сигнала, средства передачи сигнала по узкополосному каналу передачи данных. Между источником и контроллером, и между контроллером и средствами передачи сигнала расположены буферные запоминающие устройства, предназначенные для компенсации нестабильности временных интервалов обработки данных при сжатии речи.

Система функционирует следующим образом [1, 2]. Данные с аудиосенсора преобразуются в последовательность цифровых кодов и поступают на вход первого буфера для дальнейшей передачи их на обработку в ЭВМ фон-неймановской архитектуры. ЭВМ опрашивает буфер 1, используя процедуры полинга, сжимает данные программно и выдает на вход буфера 2, связывающего ЭВМ с аппаратурой передачи данных. Аппаратура передачи данных опрашивает выход буфера 2 и передает данные в канал связи, т.е. в системе присутствуют три субъекта:

1) источник аудиоданных с жесткой синхронизацией, частота которой для исключения потерь информации должна удовлетворять требования теоремы Котельникова;

2) ЭВМ фон-неймановской архитектуры, генерирующая сигналы транзакций опроса буферов и выполняющая операцию собственно сжатия данных;

3) аппаратура передачи данных с жесткой синхронизацией, частота которой должна удовлетворять стандарту интерфейса аппаратуры передачи данных с внешними устройствами.

Для оценивания времени прохождения сигнала по тракту сжатия данных и объемов буферов 1, 2 необходима математическая модель, учитывающая особенности поступления и вывода данных, и такие свойства алгоритма обработки данных как цикличность, квазистохастичность продолжения в местах ветвления и квази-случайность времени выполнения [3, 4]. Подобная модель представляет собой 3-параллельный полумарковский процесс (случайный процесс с конечным или

* Статья подготовлена в рамках выполнения Госзадания 2.3121.2017/4.6

счетным множеством состояний, имеющий ступенчатые траектории со скачками, значения в которых образуют марковский процесс) вида

$$\mu = \left\{ {}^1\mu,\ {}^2\mu,\ {}^3\mu \right\}; \qquad (1)$$

$${}^i\mu = \left\{ {}^iA,\ {}^i\mathbf{h}(t) \right\}, i = 1, 2, 3, \qquad (2)$$

где ${}^1\mu$ – процесс, описывающий ввод данных в буфер 1; ${}^2\mu$ – процесс, описывающий обработку данных, ${}^3\mu$ – процесс, описывающий вывод данных из буфера 2; ${}^iA$ – множество состояний; ${}^i\mathbf{h}(t)$ – полумарковская матрица; t – время.

В общем случае полумарковская матрица ${}^i\mathbf{h}(t)$ определена на множестве состояний ${}^iA = \left\{ {}^ia_1, ..., {}^ia_m, ..., {}^ia_M \right\}$ и имеет вид:

$${}^i\mathbf{h}(t) = \left[{}^ih_{mn}(t) \right] = \left({}^ip_{mn} \right) \otimes \left[{}^if_{mn}(t) \right],$$

где ${}^ih_{mn}(t)$ – взвешенная плотность распределения времени пребывания i-го полумарковского процесса в состоянии ${}^ia_m \in {}^iA$ с последующим переключением в состояние ${}^ia_n \in {}^iA$; ${}^ip_{mn}(t)$ – вероятность переключения i-го полумарковского процесса из ${}^ia_m$ в ${}^ia_n$; ${}^if_{mn}(t)$ – плотность распределения времени переключения из ${}^ia_m$ в ${}^ia_n$, если было принято решение подобного переключения; $\otimes$ – знак прямого перемножения матриц.

Примем, что полумарковские процессы (1) являются эргодическими. В самом общем случае структуры подобных процессов представляют собой полные графы с петлями, приведенные на рис. 1b, 1c, 1d.

Процессы ${}^1\mu$ и ${}^3\mu$ (рис. 1b, 1d) имеют вид

$${}^i\mu = \left\{ \left\{ {}^ia \right\}, {}^ih(t) \right\},\ i = 1, 3, \qquad (3)$$

где $\left\{ {}^ia \right\}$ – множество, состоящее из единственного состояния генерации транзакций; ${}^ih(t) = \delta\left(t - {}^iT\right)$; $\delta(...)$ – δ-функция Дирака; ${}^iT$ – период следования транзакций; $i = 1, 3$ (рис. 1b, 1d).

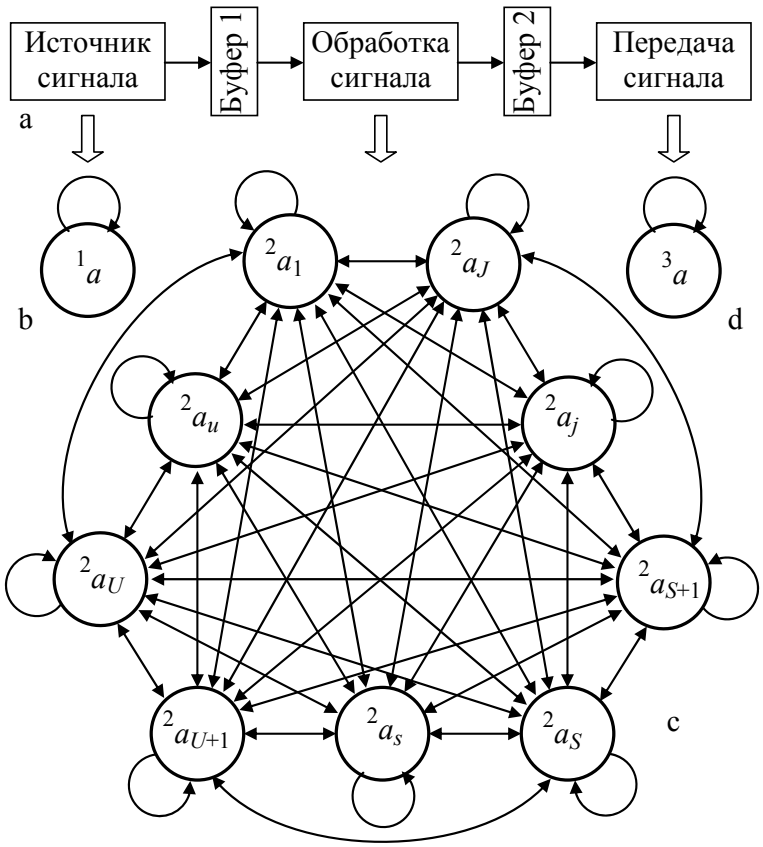

Рис. 1. Структура системы передачи данных со сжатием (a)
и моделей генераторов транзакций – источника (b), обработчика (c), передатчика (d)

Структура полумарковского процесса ${}^2\mu$ приведена на рис. 1 с. В указанном процессе

$$
\begin{aligned}
&{}^2\mathbf{h}(t)=\left[{}^2h_{mn}(t)\right];\\
&{}^2A={}^2A_u\cup{}^2A_s\cup{}^2A_j,
\end{aligned}
\qquad (4)
$$

где ${}^2\mathbf{h}(t)$ имеет размер $J\times J$; $h_{mn}(t)$ – взвешенные плотности распределения;

${}^2A_u=\left\{{}^2a_1,...,{}^2a_u,...,{}^2a_U\right\}$ – подмножество состояний генерации запросов в буфер 1;

${}^2A_s=\left\{{}^2a_{U+1},...,{}^2a_s,...,{}^2a_S\right\}$ – подмножество состояний генерации запросов в буфер 2;

${}^2A_J=\left\{{}^2a_{S+1},...,{}^2a_j,...,{}^2a_J\right\}$ – подмножество состояний, моделирующих прочие операторы алгоритма; $\left|{}^2A\right|=J$.

В полумарковских процессах ${}^1\mu$, ${}^3\mu$ транзакции генерируются при каждом переключении процесса из состояния ${}^ia$ в состояние ${}^ia$, $i=1,3$. В полумарковском процессе ${}^2\mu$ транзакция генерируется в одном из двух случаев:

1) при прямом переключении процесса из состояний с номерами с 1 по S в состояния с теми же номерами;

2) при переключении процесса из состояний с номерами с 1 по S в состояния с номерами с $S+1$ по J с последующим блужданием из состояний с номерами с $S+1$ по J в состояния с номерами с 1 по S.

Методами, приведенными в [5, 6], полумарковский процесс ${}^2\mu$ может быть упрощен до процесса ${}^2\tilde{\mu}$, включающего только состояния генерации транзакций:

$$
{}^2\mu\rightarrow{}^2\tilde{\mu}=\left\{{}^2\tilde{A},\,{}^2\tilde{\mathbf{h}}\{t\}\right\},
\qquad (5)
$$

где ${}^2\tilde{A}=\left\{{}^2\tilde{a}_1,...,{}^2\tilde{a}_u,...,{}^2\tilde{a}_U,\,{}^2\tilde{a}_{U+1},...,{}^2\tilde{a}_s,...,{}^2\tilde{a}_S\right\}$ – сокращенное множество вершин; ${}^2\tilde{\mathbf{h}}(t)=\left[\tilde{h}_{mn}(t)\right]$ – полумарковская матрица размером $S\times S$; $1\le m,n\le S$.

Процесс ${}^2\tilde{\mu}$ остается эргодическим. При каждом переключении полумарковского процесса ${}^2\tilde{\mu}$ генерируется одна транзакция, либо в буфер 1, либо в буфер 2. Для внешнего наблюдателя вероятности пребывания в состоянии m в установившемся режиме переключений определяются по зависимостям:

$$
\pi_m=\frac{T_m}{\theta_m},
\qquad (6)
$$

где T_m – математическое ожидание времени пребывания процесса ${}^2\tilde{\mu}$ в состоянии ${}^2\tilde{a}_m\in{}^2\tilde{A}$; θ_m – время возврата в состояние ${}^2\tilde{a}_m\in{}^2\tilde{A}$.

Время T_m определяется по зависимости

$$
T_m=\int_0^\infty t\cdot\sum_{m=1}^{S}\tilde{h}_{mn}(t)dt.
\qquad (7)
$$

Для определения θ_m расщепим ${}^2\tilde{a}_m$ на ${}^{2b}\tilde{a}_m$ и ${}^{2e}\tilde{a}_m$. Это осуществляется за счет переноса столбца матрицы ${}^2\tilde{\mathbf{h}}(t)$ с номером m в столбец с номером $S+1$. Столбец с номером m и строка с номером $S+1$ заполняются нулями. В результате этого формируется матрица ${}^2\tilde{\mathbf{h}}'(t)$, имеющая размерность $(S+1)\times(S+1)$, Математическое ожидание времени возврата определяется по следующей зависимости:

$$
\theta_m=\int_0^\infty t\cdot L^{-1}{}^r\mathbf{I}_{S+1}\cdot\sum_{k=1}^{\infty}\left\{L\left[\tilde{\mathbf{h}}'(t)\right]\right\}^k\cdot{}^c\mathbf{I}_m dt,
\qquad (8)
$$

где ${}^c\mathbf{I}_m$ – вектор-столбец, имеющий размер $S+1$, m-й элемент которого равен единице, а остальные элементы равны нулю; ${}^r\mathbf{I}_m$ – вектор-строка, имеющий размер $S+1$, $(S+1)$-й элемент которого равен единице, а остальные элементы равны нулю; $L[...], L^{-1}[...]$ – прямое и обратное преобразования Лапласа.

С учетом (6) и свойства эргодичности полумарковского процесса, плотность распределения времени между двумя транзакциями в буфер 1 будет равна

$$
f_{UU}(t)=\frac{\sum\limits_{m=1}^{U}\pi_m\sum\limits_{n=1}^{U}\tilde{h}_{mn}(t)}{\sum\limits_{m=1}^{U}\pi_{m_3}\sum\limits_{n=1}^{U}\tilde{p}_{mn}};
\qquad (9)
$$

плотность распределения времени между двумя транзакциями в буфер 2 будет равна

$$
f_{SS}(t)=\frac{\sum\limits_{m=U+1}^{S}\pi_m\sum\limits_{n=U+1}^{S}\tilde{h}_{mn}(t)}{\sum\limits_{m=U+1}^{S}\pi_{m_3}\sum\limits_{n=U+1}^{S}\tilde{p}_{mn}};
\qquad (10)
$$

плотность распределения времени между транзакцией в буфер 1 и транзакцией в буфер 2

$$
f_{US}(t)=\frac{\sum\limits_{m=1}^{U}\pi_m\sum\limits_{n=U+1}^{S}\tilde{h}_{mn}(t)}{\sum\limits_{m=1}^{U}\pi_{m_3}\sum\limits_{n=U+1}^{S}\tilde{p}_{mn}};
\qquad (11)
$$

плотность распределения времени между транзакцией в буфер 2 и транзакцией в буфер 1

$$f_{SU}(t)=\frac{\sum_{m=U+1}^{S}\pi_m\sum_{n=1}^{U}\tilde{h}_{mn}(t)}{\sum_{m=U+1}^{S}\pi_{m_3}\sum_{n=1}^{U}\tilde{p}_{mn}}. \quad (12)$$

В формулах (9) – (12)

$$\tilde{p}_{mn}=\int_0^{\infty}\tilde{h}_{mn}(t)dt). \quad (13)$$

Вследствие того, что транзакции генерируются в результате блужданий по состояниям полумарковских процессов, транзакции, сгенерированные по каждой отдельной траектории, могут рассматриваться как отдельный поток, а генерация по множеству возможных траекторий может рассматриваться как объединение потоков транзакций. В соответствии с теоремой Б. Григелиониса [7], подобный суммарный поток является пуассоновским. Следовательно, можно ввести ограничение на плотности распределения времени между транзакциями, и считать, что процесс является строго марковским с непрерывным временем [8], а указанные плотности описываются следующим образом:

$${}^{2}f_i(t)=\frac{1}{{}^{2}T_i}\exp\left(-\frac{t}{{}^{2}T_i}\right),\ i=1,3, \quad (14)$$

где

${}^{2}f_1(t)=f_{UU}(t);\ {}^{2}f_3(t)=f_{SS}(t);\ {}^{2}T_i=\int_0^{\infty}t\cdot{}^{2}f_i(t)dt,$

$i=1,3$ – математическое ожидание времени между двумя последовательными транзакциями в соответствующем потоке.

«СОРЕВНОВАНИЕ» ЗА ЗАПОЛНЕНИЕ/ОПОРОЖНЕНИЕ БУФЕРА

Для оценки объемов буферов сделаем допущение о том, что полумарковские процессы ${}^{1}\mu$ и ${}^{3}\mu$ также являются строго марковскими с непрерывным временем, и потоки транзакций при заполнении буфера 1 и опорожнении буфера 2 являются пуассоновскими, т.е.

$${}^{i}h(t)=\frac{1}{{}^{i}T}\exp\left(-\frac{t}{{}^{i}T}\right),\ i=1,3. \quad (15)$$

Для этого случая мажет быть разработана математическая модель «соревнования» [9, 10, 11] за заполнение и опорожнение абстрактного буфера, которая представляет собой цепочку 2-параллельных марковских процессов вида (рис. 2 а):

$${}^{c}\mu=\left\{{}^{c}\mathrm{B},\ {}^{c}\boldsymbol{\eta}(t)\right\}, \quad (16)$$

где ${}^{c}\mathrm{B}=\left\{{}^{c}\mathrm{B}_0,...,{}^{c}\mathrm{B}_n,...,{}^{c}\mathrm{B}_N\right\}$ – множество состояний; ${}^{c}\boldsymbol{\eta}(t)$ – марковская матрица.

Каждое состояние цепочки (16) представляет собой 2-параллельный марковский процесс вида

$${}^{c}\mathrm{B}_n=\left\{\left\{{}^{c}\beta_{1b}^{n},\ {}^{c}\beta_{1e}^{n},\ {}^{c}\beta_{2b}^{n},\ {}^{c}\beta_{2e}^{n}\right\},\ {}^{c}\boldsymbol{\eta}_n(t)\right\},\quad 0\le n\le N, \quad (17)$$

где

$${}^{c}\boldsymbol{\eta}_n(t)=\begin{bmatrix}\begin{bmatrix}0 & \frac{1}{\tau_+}\exp\left(-\frac{t}{\tau_+}\right)\\ 0 & 0\end{bmatrix} & \mathbf{0}\\ \mathbf{0} & \begin{bmatrix}0 & \frac{1}{\tau_-}\exp\left(-\frac{t}{\tau_-}\right)\\ 0 & 0\end{bmatrix}\end{bmatrix},$$

$0\le n\le N.$ (18)

τ_+ – математическое ожидание времени между двумя транзакциями по заполнению буфера; τ_- – математическое ожидание времени между двумя транзакциями по опорожнению буфера.

Если привязать параметры марковского процесса (18) к параметрам процессов (1), то

$$\tau_+=\begin{cases}{}^{1}T, if\ \text{buffer 1},\\ {}^{2}T_3, if\ \text{buffer 2};\end{cases} \quad (19)$$

$$\tau_-=\begin{cases}{}^{2}T_1, if\ \text{buffer 1},\\ {}^{3}T, if\ \text{buffer 2}.\end{cases} \quad (20)$$

В свою очередь, цепочка из 2-параллельных марковских процессов может быть преобразована в ординарный марковский процесс

$$\tilde{\mu}=\left\{\mathrm{B},\boldsymbol{\eta}(t)\right\}, \quad (21)$$

где $\mathrm{B}=\left\{\beta_0,...,\beta_n,...,\beta_N\right\}$ – множество состояний, каждое из которых определяется количеством занятых в текущий момент ячеек буфера; $\boldsymbol{\eta}(t)=\left[\eta_{mn}(t)\right]$ – марковская матрица размером $(N+1)\times(N+1)$, у которой

$$\eta_{mn}(t)=\begin{cases}\frac{1}{\tau}\exp\left(-t\frac{\alpha+1}{\tau}\right), if\ n=m+1;\\ 0, if\ n\ne m+1, \text{or}\ n\ne m-1;\\ \frac{\alpha}{\tau}\exp\left(-t\frac{\alpha+1}{\tau}\right), if\ n=m-1;\end{cases} \quad (22)$$

$$\alpha=\frac{\tau_+}{\tau_-}. \quad (23)$$

Структура марковского процесса показана на рис. 2b.

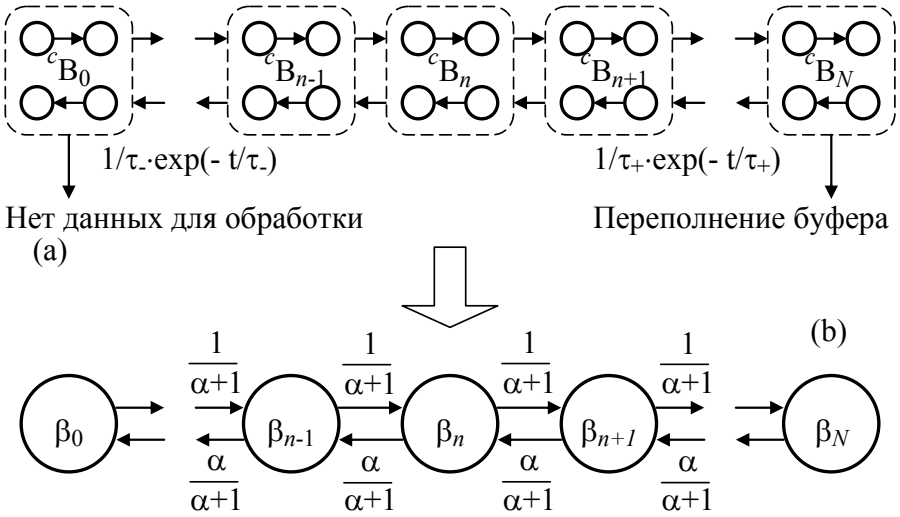

Рис. 2. Марковский процесс, описывающий «соревнование» за заполнение и опорожнение абстрактного буфера (a) и расчетная модель (b)

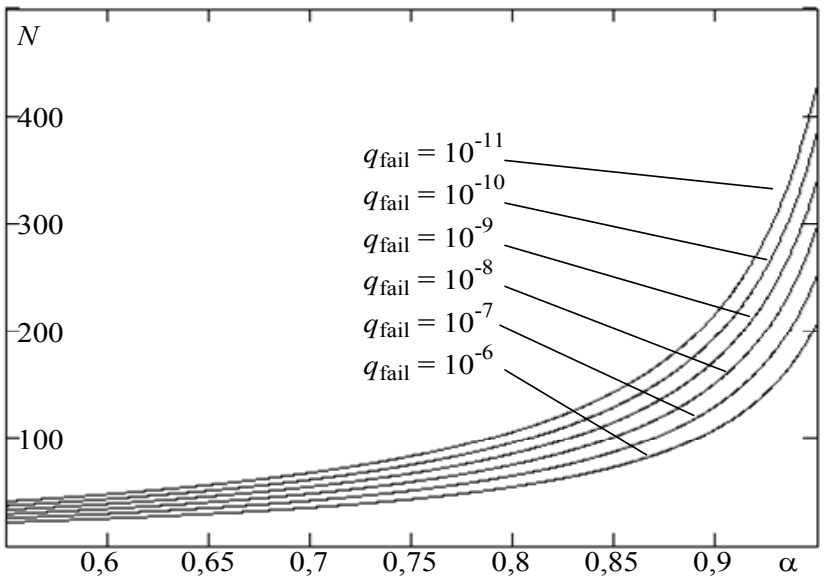

Рис. 3. Зависимость потребного объема буфера от соотношения периодов поступления и считывания данных и от допустимой вероятности сбоя

ОЦЕНКА ПОТРЕБНОГО ОБЪЕМА БУФЕРА

Для определения потребного объема абстрактного буфера может быть применена теория массового обслуживания [12], однако упростить расчеты позволяет зависимость (6), примененная к марковскому процессу $\tilde{\mu}$. Расчет оценок длительностей временных интервалов и вероятностей осуществим по зависимостям, приведенным в [5, 6]. Вероятности того, что для внешнего наблюдателя в произвольный момент времени процесс (16) будет находиться в состояниях β_n, будут равны

$$q_n = \frac{\alpha^n}{\sum_{i=0}^{N} \alpha^i}. \qquad (24)$$

Сбой работы системы из-за переполненного буфера наступает при заполнении всех ячеек буфера, т.е.

$$q_{fail} = \frac{\alpha^N}{\sum_{i=0}^{N} \alpha^i}. \qquad (25)$$

Умножив обе части (22) на $\alpha - 1$, будем иметь

$$q_{\text{fail}} = \frac{\alpha^N (\alpha - 1)}{\alpha^{N+1} - 1}. \qquad (26)$$

Из (23) следует, что

$$N \geq \frac{\ln q_{fail} - \ln\left[1-\alpha\left(1-q_{\text{fail}}\right)\right]}{\ln\alpha}, \qquad (27)$$

где α – соотношение между математическими ожиданиями плотностей распределения времени заполнения и опорожнения абстрактного буфера; q_{fail} – допустимая вероятность сбоя из-за переполнения буфера; N – объем буфера.

В качестве примера рассмотрим случай, когда $0{,}55 \leq \alpha \leq 0{,}95$. Графики зависимости $N(\alpha)$ для различных q_{fail} приведены на рис. 3.

ЗАКЛЮЧЕНИЕ

В результате выполненного нами исследования разработана методика оценивания временных интервалов между транзакциями в алгоритме сжатия речевых сообщений, позволяющая рассчитать объем буфера между источником и приемником, обеспечивающего при обработке непрерывных потоков аудиоданных вероятность отказа не ниже априорно заданной.

Результаты исследования могут быть использованы в интересах оптимизации аппаратного и программного обеспечения систем сжатия и для расчетов объемов буферов для конкретных алгоритмов сжатия данных.

СПИСОК ЛИТЕРАТУРЫ

1. Ramírez M.A., Minami M. Technology and standards for low-bit-rate vocoding methods // The Handbook of Computer Networks / ed. H. Bidgoli. – 2011. Vol. 2. – P. 447 – 467.
2. NATO Interoperable Narrow Band Voice Coder. STANAG 4591 C3 (Edition 1) – The 600 bit/s, 1200 bit/s, 2400 bit/s. NSA/1025(2008)-C3/4591.
3. Korolyuk V., Swishchuk A. Semi-Markov random evolutions // Semi-Markov Random Evolutions. – Springer Science + Business Media, 1995. – P. 59–91.
4. Ивутин А.Н., Ларкин Е.В. Обобщенная полумарковская модель алгоритма управления цифровыми устройствами // Известия Тульского государственного университета. Технические науки. – 2013. – № 1. – С. 221-227.
5. Ивутин А.Н., Ларкин Е.В., Костомаров Д.С. Методика формирования сети Петри-Маркова для моделирования когнитивных технологий // Известия Тульского государственного университета. Технические науки. – 2013. – № 9-1. – С. 303-310.
6. Есиков Д.О., Ларкин Е.В., Ивутин А.Н. Математические модели и алгоритмы решения задачи обеспечения устойчивости функционирования распределенных информационных систем // Математические методы в технике и технологиях – ММТТ. – 2016. – № 2 (84). – С. 133-137.
7. Григелионис Б. О сходимости сумм ступенчатых случайных процессов к пуассоновскому // Теория вероятностей и ее применение. – 1963. – Т. 8, №2. – С. 189-194.
8. Марков А.А. Распространение закона больших чисел на величины, зависящие друг от друга // Известия физико-математического общества при Казанском университете. Сер. 2. – 1906. – Том 15. – С. 135-156.
9. Cleaveland R., Smolka S.A. Strategic directions in concurrency research // CSUR. –1996. – Vol. 28, № 4. – P. 607 – 625.
10. Ivutin A.N, Larkin E.V. Simulation of Concurrent Games // Вестник Южно-Уральского государственного университета. Серия: Математическое моделирование и программирование. – 2015. – Т. 8, № 2. – С. 43-54.
11. Larkin E.V., Ivutin A.N., Kotov V.V., Privalov A.N. Simulation of Relay-races // Вестник Южно-Уральского государственного университета. Серия: Математическое моделирование и программирование. – 2016. – Т. 9, № 4. – С. 117-128.
12. Sundarapandian V. Queueing Theory: Probability, Statistics and Queueing Theory. –New Delhi: PHI Learning, 2009. – 144 p.

Материал поступил в редакцию 29.04.17

Сведения об авторах

ЛАРКИН Евгений Васильевич – доктор технических наук, профессор, заведующий кафедрой робототехники и автоматизации производства ФГБОУ ВПО «Тульский государственный университет»
e-mail: elarkin@mail.ru

БОГОМОЛОВ Алексей Валерьевич – доктор технических наук, профессор, ведущий научный сотрудник, ФГКВОУ ВО «Военная академия Генерального штаба Вооруженных Сил Российской Федерации»,
e-mail: a.v.bogomolov@gmail.com

ПРИВАЛОВ Александр Николаевич – доктор технических наук, профессор, заведующий кафедрой информационных технологий ФГБОУ ВПО «Тульский государственный педагогический университет им. Льва Толстого»,
e-mail: privalov.61@mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322

В.Н. Захаров, Ю. В. Никитин, Ал-др А. Хорошилов, А. А. Хорошилов

Технологии создания новых направлений перевода для системы метафраз (на примере казахско-русского перевода)

Рассматривается технология создания новых направлений перевода для системы МетаФраз (на примере казахско-русского перевода): проведена работа по формализации грамматики казахского языка; разработаны грамматические признаки казахских слов, и на основе метода лингвистической аналогии созданы машинные грамматические таблицы для их морфологического анализа. При формировании двуязычных машинных словарей использовались технологии создания словарей по параллельным текстам (билингвам); разработана система правил для семантико-синтаксического анализа текстов на казахском языке. Эти правила были интегрированы в имеющиеся процедуры семантико-синтаксического анализа и трансфера программно-лингвистической платформы Metafraz. На основе проведенных исследований создано программное обеспечение, осуществляющее процесс автоматического и интерактивного перевода текстов с казахского на русский язык.

Ключевые слова: *автоматический перевод текстов, казахско-русский перевод, формализованное описание текста, смысловая структура, лингвистическое программное обеспечение, декларативные средства*

ВВЕДЕНИЕ

Современные системы машинного перевода с одних естественных языков на другие можно условно разделить на три категории: системы, функционирующие на основе грамматических правил (Rule-Based Machine Translation – RBMT), статистические системы (Statistical Machine Translation – SMT) и гибридные системы, сочетающие преимущества и тех, и других. Системы RBMT анализируют текст и строят его перевод на базе встроенных словарей и набора правил для данной языковой пары. Системы SMT действуют совсем по иному принципу: в их основе лежат математические методы для получения перевода. Точнее, весь принцип работы подобной системы основан на статистическом вычислении вероятности совпадений фраз из исходного текста с фразами, которые хранятся в базе системы перевода. К гибридным системам можно отнести системы фразеологического машинного перевода (Phraseological Machine Translation – FMT), которые реализуют все возможности традиционных систем RBMT, но при этом широко используют принцип лингвистической аналогии при разработке грамматических правил и словарной системы. Схожесть подходов систем SMT и FMT заключается в том, что в их основе лежит идея – опираться при машинном переводе на ранее выполненные переводы текстов, представленные в виде *больших объемов языковых пар.* Различием между ними является то обстоятельство, что при реализации этого принципа в технологиях FMT обеспечивается возможность, используя методы статистического установления переводных соответствий для слов и словосочетаний, составлять адекватные статистически обоснованные двуязычные частотные словари по массивам двуязычных текстов. При этом технологии FMT полностью исключают фатальную зависимость создания декларативных средств (как в случае подходов SMT) от наличия требуемых объемов параллельных текстов [1, 2].

ПРИНЦИПЫ ПОСТРОЕНИЯ СИСТЕМ ФРАЗЕОЛОГИЧЕСКОГО МАШИННОГО ПЕРЕВОДА

В соответствии с концепцией, сформулированной известным российским ученым проф. Г.Г. Белоноговым [1, 3], система фразеологического машинного перевода должна включать в свой состав компьютерную словарную базу, содержащую переводные эквиваленты не только для отдельных слов, но и для часто встречающихся словосочетаний и фраз. В процессе перевода текстов система должна использовать хранящиеся в этой

базе переводные эквиваленты в следующем порядке: сначала для очередного предложения исходного текста делается попытка перевести его как целостную фразеологическую единицу; далее, в случае неудачи, – переводятся входящие в состав текста словосочетания; и, наконец, осуществляется пословный перевод тех фрагментов предложения, которые не удалось перевести первыми двумя способами. Подобный подход к процессу автоматизированного перевода позволяет, наряду с традиционным переводом, использовать возможности статистического перевода и подхода Translation Memory (TM).

Подход TM заключается в том, что в процессе перевода накапливаются и заносятся в память крупные двуязычные текстовые блоки (например, целые предложения), которые используются при дальнейшем переводе.

Процесс перевода текстов с одного естественного языка на другой в системах FMT включает три основных этапа: семантико-синтаксический анализ исходного текста, трансфер и семантико-синтаксический синтез выходного текста. Обобщенная схема процесса перевода текстов в системах FMT приведена в табл.1.

В соответствии с предлагаемой нами технологией создания «нового» направления перевода, необходимо для этого «нового» языка по единой методике разработать комплекс декларативных и процедурных средств морфологического, семантико-синтаксического и концептуального анализа и синтеза текстов. Для этого необходимо:

1) разработать общие принципы построения и формализации машинной грамматики казахского языка на основе теоретической концепции фразеологического машинного перевода;

2) решить задачи кодировки алфавита казахского языка;

3) составить частотные словари слов и представить их в прямом и обратном порядке по репрезентативным корпусам казахских текстов;

4) разработать методы, программы и технологии составления двуязычных словарей по параллельным текстам;

5) составить двуязычные словари слов и словосочетаний;

6) разработать методы и технологии формирования грамматических таблиц для морфологического анализа казахских текстов;

7) разработать процедуру графематического анализа казахских текстов;

8) разработать процедуру морфологического анализа казахских текстов;

9) разработать процедуру лемматизации слов казахских текстов;

10) разработать процедуру упрощенного семантико-синтаксического анализа казахских текстов;

11) разработать процедуры упрощенного концептуального анализа казахских текстов;

12) разработать служебные функции, обеспечивающие процесс казахско-русского перевода и построение формализованного представления казахских текстов;

13) реализовать процесс автоматического перевода казахских текстов;

14) реализовать процесс интерактивного перевода казахских текстов;

15) разработать пользовательские и служебные интерфейсы, обеспечивающие режимы перевода казахских текстов;

16) разработать методы, программы и технологие формирования машинных двуязычных казахско-русских словарей;

17) разработать макет системы казахско-русского перевода;

18) выполнить комплексную отладку макета системы казахско-русского перевода.

Таблица 1

Обобщенная схема процесса перевода текстов в системах FMT

№№ п/п	Наименование этапов перевода
1	**Семантико-синтаксический анализ исходного текста на казахском языке:** 1.1.Членение текста на исходные предложения 1.2. Морфологический анализ слов исходного текста 1.3. Семантико-синтаксический (концептуальный и синтаксический) анализ исходного текста
2	**Трансфер:** 2.1. Замена наименований понятий (слов и словосочетаний) исходного текста на наименования понятий выходного текста 2.2. Преобразование информации о синтаксической структуре исходного текста в информацию, необходимую для синтеза выходного текста (в частности, присвоение словам и словосочетаниям грамматических признаков членов предложения и установление между ними отношений непосредственной доминации)
3	**Семантико-синтаксический синтез выходного русского текста:** 3.1. Морфологический анализ слов переводных соответствий с целью определения их грамматических признаков – признаков их грамматических классов и признаков формы (род, число, падеж, лицо и др.) 3.2. Формирование синтаксической структуры выходного текста на основе результатов выполнения п. 2.2 3.3. Морфологический синтез форм исходных слов в соответствии с их словоизменительными моделями и грамматической ролью в предложении

Следует отметить, что решение этой задачи значительно облегчается тем, что в рамках программно-лингвистической платформы МетаФраз [1, 3] ранее были созданы англо-русское и немецко-русское направления перевода. Поэтому для поставленной нами задачи не было необходимости разрабатывать процедуры трансфера и семантико-синтаксического синтеза русских текстов. В связи с этим наши основные усилия были направлены на разработку процедур морфологического, семантико-синтаксического и концептуального анализа текстов на казахском языке, а также средств создания фразеологических казахско-русских словарей.

ОБЩАЯ ХАРАКТЕРИСТИКА КАЗАХСКОГО ЯЗЫКА

Казахский язык – это национальный язык коренного населения Республики Казахстан и проживающих за ее пределами представителей казахского народа (Россия, Узбекистан, Китай, Монголия и др.). Он является государственным языком Республики Казахстан, входит в кыпчакскую группу (северо-восточный ареал) тюркских языков (татарский, башкирский, карачаево-балкарский, кумыкский, караимский, крымско-татарский каракалпакский, ногайский) и наиболее близок к ногайскому и каракалпакскому языкам. Казахская письменность претерпела ряд изменений. Так, до 1929 г. в ее основе лежала арабская графика, в 1929–1940 гг. – латинская графика, а с 1940 г. по настоящее время – русская графика (кириллица).

Казахский кириллический алфавит используется в Казахстане и Монголии. Этот алфавит, разработанный С.А. Аманжоловым и принятый в 1940 г., содержит 42 буквы: 33 буквы русского алфавита и 9 специфических букв казахского языка. Вначале казахские буквы размещались после букв русского алфавита, затем были перенесены на места после русских букв, сходных по звучанию.

Казахский язык – агглютинативный язык. Это означает, что словоизменение происходит последовательным присоединением к неизменной основе аффиксов (приставки в казахском языке отсутствуют); существительное не имеет категории рода, зато имеет категорию принадлежности; прилагательные не согласуются ни в числе, ни в падеже; имеет шесть (вместе с формальной формой восемь) личных местоимений.

Еще одна особенность казахского языка состоит в отсутствии предлогов. Значение русских предлогов большей частью передается посредством послелогов, а также в форме косвенных падежей. Если перед существительным стоит числительное, то окончание множественного числа в существительном не употребляется. Числительные и прилагательные в функции определения перед существительными не изменяются ни в числе, ни в падеже.

В казахском языке отсутствует категория рода, поэтому одно и то же прилагательное, местоимение или порядковое числительное, в зависимости от смысла предложения, может переводиться на русский язык в мужском, женском или среднем роде. Именные части речи, в отличие от русского языка, изменяются по лицам.

Причастие, стоящее перед определяемым существительным, не изменяется ни по падежам, ни по числам в отличие от русского языка, где причастие согласуется с определяемым словом в роде, числе, падеже. Имена числительные (количественные и порядковые), употребляясь в предложении в функции определения, не изменяются в числе и в падеже.

СОЗДАНИЕ МАШИННОЙ ГРАММАТИКИ ДЛЯ КАЗАХСКОГО ЯЗЫКА

Для процедур семантико-синтаксического анализа необходимо, чтобы морфологический анализ правильно устанавливал морфологическую структуру слов и назначал им грамматические признаки. При этом и морфологическая структура казахских слов, и их грамматические признаки должны соотноситься с системой категоризации русских слов, которая была положена в основу разработки таблицы грамматических признаков для 14-ти лексико-грамматических классов слов (табл. 2).

Таблица 2

Лексико-грамматические классы слов и их мнемоническое обозначение

	Грамматический класс слова	Мнемоническое обозначение
1	Существительное	N
2	Прилагательное	A
3	Глагол	V
4	Субстантивированное прилагательное	S
5	Местоимение	M
6	Числительное	C
7	Союз	&
8	Междометие	!
9	Наречие	Y
10	Предлог	F
11	Послелог	B
12	Причастие	W
13	Деепричастие	D
14	Вводное слово	P

Таблица 3

Фрагмент таблицы грамматических признаков для слов казахского языка

Существительное (N)				
I	***Тип лексико-грамматического класса***	***Мнемоника***	***Местоположение в ГИ***	
			группа	***позиция***
	Существительное	*N*	*2*	*1*
II	***Тип существительного***	***Мнемоника***	***Местоположение в ГИ***	
			группа	***позиция***
0	*Не определено*	*0*	*2*	*2*
1	*Нарицательное*	*N*	*2*	*2*
2	*Фамильно-именная группа*	*F*	*2*	*2*
6	*Топоним*	*T*	*2*	*2*
7	*Имя собственное*	*F*	*2*	*2*
8	*Название фирмы*	*B*	*2*	*2*
9	*Уменьшительное*	*L*	*2*	*2*
10	*Вспомогательное*	*A*	*2*	*2*

III	***Категория числа***	***Мнемоника***	***Местоположение в ГИ***	
			группа	***позиция***
0	*Не определено*	*0*	*3*	*1*
1	*Ед. число*	*e*	*3*	*1*
2	*Мн.число*	*m*	*3*	*1*

IV.	***Категория падежа***	***Мнемоника***	***Местоположение в ГИ***	
			группа	***позиция***
0	*Не определено*	*0*	*3*	*2*
1	*Именительный*	*1*	*3*	*2*
2	*Родительный*	*2*	*3*	*2*
3	*Дательно-направительный*	*3*	*3*	*2*
4	*Винительный*	*4*	*3*	*2*
5	*Творительный*	*5*	*3*	*2*
6	*Исходный*	*a*	*3*	*2*
7	*Местный*	*m*	*3*	*2*
8	*Местно-наречный*	*y*	*3*	*2*

Фрагмент таблицы грамматических признаков для слов казахского языка представлен в табл. 3, где для каждого признака указано его местоположение в позиционном представлении набора признаков. Грамматические признаки слов разделены на три группы: характеризирующие структуру слова (длину изменяемой части в пределах словоизменительной парадигмы), обозначающие лексико-грамматический класс и его подтип, а также дополнительные грамматические признаки, связанные с формой представления данного слова (падеж, число, принадлежность, лицо и др.). Каждый признак обозначен буквенно-цифровой мнемоникой. Местоположение значения признака в наборе грамматической информации (ГИ) указано номером группы и его позицией в этой группе. Разработанный набор грамматических признаков таблицы позволяет решить задачу лемматизации казахских слов путем присоединения к основе нормализующих окончаний.

ДЕКЛАРАТИВНЫЕ СРЕДСТВА МАШИННОЙ ГРАММАТИКИ КАЗАХСКИХ СЛОВ

Чтобы решить задачу автоматизации, корректировки и ввода основной и дополнительной грамматической информации для слов казахского языка были разработаны автоматизированные средства поддержки этих процессов. В качестве исходного массива для назначения грамматической информации был выбран частотный словарь слов, полученный по массиву казахских текстов (объемом более 400 тыс. слов); его объем составил около 50 тыс. слов. Для этого частотный словарь предварительно упорядочивался в обратном лексико-грамматическом порядке и всем словам словаря автоматически устанавливалась грамматическая информация по списку суффиксов (или сочетаний суффиксов) и окончаний.

Таблица 4

Фрагменты обратного словаря казахских слов с назначенной грамматической информацией

.....абайлаттырғанбысыз	00/VW/0iu2
.....абаласқанбысыз	00/VW/0is2
.....абайласқанбысыз	00/VW/0is2
.....абаландырысқанбысыз	00/VW/0i32
.....абайландырысқанбысыз	00/VW/0i32
.....абалатысқанбысыз	00/VW/0i22
.....абайлатысқанбысыз	00/VW/0i22
.....абалатқанбысыз	00/VW/0iu2
.....абайлатқанбысыз	00/VW/0iu2
.....	.....
.....абалалармысыңдар	00/VW/0ip3
.....абайлалармысыңдар	00/VW/0ip3
.....абайлармысыңдар	00/VW/0ia3
.....абаландырармысыңдар	00/VW/0i43
.....абайландырармысыңдар	00/VW/0i43
.....абалаттырармысыңдар	00/VW/0iu3

Для такого анализа всем суффиксам и окончаниям ставились в соответствие грамматические признаки, которые могли бы быть совместимы с ними. При этом допускалась многозначность назначения грамматических признаков. Корректировку и разрешение многозначности выполнял человек-лингвист с помощью специального пользовательского интерфейса.

В процессе автоматизированной проверки и корректировки грамматической информации для казахских слов массива, упорядоченного в обратном порядке, лингвист, в случаях неправильного назначения грамматической информации, вносил изменения в соответствующие окна пользовательского интерфейса путем выбора требуемых пунктов выпадающего меню для каждого грамматического признака. Фрагменты обратного словаря казахских слов с назначенной грамматической информацией приведены в табл. 4.

Мнемонические обозначения грамматических признаков указываются в соответствии с табл. 2. Полученный словарь положен в основу двух таблиц для морфологического анализа казахских слов – таблицы конечных буквосочетаний слов (КБС) и таблицы коротких и служебных слов (СКС). Технология создания таких таблиц подробно изложена в работах [1, 3].

Таблица КБС, предназначенная для обработки слов по методу аналогии, включает 8457 конечных буквосочетаний; таблица СКС содержит следующие классы слов: местоимения, послелоги, союзы и короткие слова, буквенный состав которых менее пяти букв. Объем полученного словаря составляет 1342 слова.

ТЕХНОЛОГИИ СОСТАВЛЕНИЯ ДВУЯЗЫЧНЫХ СЛОВАРЕЙ

В концепции фразеологического машинного перевода особое внимание уделяется автоматизированным технологиям составления двуязычных словарей [1]. Таких технологий разработано несколько, и их применение определяется конкретными задачами. Первоначальная версия казахско-русского словаря создавалась путем семантико-синтаксического и статистического анализа вышеупомянутого корпуса казахских текстов. В результате был выявлен статистически обоснованный понятийный состав этого корпуса текстов, включающий 57643 казахских слов и словосочетаний. Далее эти слова и словосочетания были вручную переведены на русский язык и включены в первоначальный состав словарной базы системы казахско-русского перевода.

На следующем этапе была использована технология составления фразеологических словарей по параллельным текстам (билингвам) [2], базирующаяся на гипотезе: *если два разноязычных предложения из массива двуязычных текстов (билингв) являются переводами друг друга, то для каждого отрезка текста (слова или словосочетания), входящего в состав предложения на одном языке, с высокой вероятностью найдется эквивалентный ему по смыслу отрезок текста, входящий в предложение на другом языке.* Следовательно, если для некоторого казахского словосочетания (наименования понятия) подобрать множество включающих его казахских предложений и множество русских переводов этих предложений, то в русских предложениях будут многократно встречаться переводы этого словосочетания.

Для решения этой задачи были разработаны процедуры: выделения наименований понятий в текстах на казахском языке; выделения наименований понятий в русских текстах и автоматического установления смысловой близости между разноязычными наименованиями понятий.

Руководствуясь предложенной нами гипотезой и располагая набором процедурных и декларативных средств, был проведен эксперимент по автоматизированному составлению казахско-русских словарей с использованием в качестве тестового массива 100 тыс. параллельных (казахских и русских) предложений, являющихся переводами друг друга.

Процесс автоматизации составления двуязычных фразеологических словарей по двуязычным параллельным текстам включает пять этапов.

1. Членение исходных казахских и русских текстов на отдельные предложения и формирование массива пар казахских и русских предложений, которые являются переводами друг друга. При этом такие пары разноязычных предложений нумеруются. Установление смысловой близости предложений выполняется с помощью процедуры автоматического распознавания их смыслового тождества или смысловой близости.

2. Концептуальный анализ казахского предложения, получение возможных наименований понятий и формирование выборок пар предложений, в которых содержатся казахские наименования понятий. Далее по русской части пар предложений составляется частотный словарь русских наименований понятий. В качестве возможных кандидатов на роль переводного эквивалента берутся самые частые словосочетания, в состав которых входят переводы слов (или их синонимов) в обоих направлениях.

3. Установление наиболее вероятных переводных соответствий между фрагментами (словами и словосочетаниями) казахского и русского предложений. При этом с целью повышения полноты установления переводных соответствий все слова русских переводных эквивалентов казахских наименований понятий и все слова русского предложения нормализуются (приводятся к словарной форме). Далее работа выполняется в следующем порядке: в массиве предложений, сформированном в п. 2, выделяется его первый казахский фрагмент (слово или словосочетание), и словарные варианты его русских переводных эквивалентов ищутся в русском предложении в направлении слева направо (от его начала до его конца). Если один из переводных эквивалентов первого казахского фрагмента совпадает с каким-либо фрагментом русского предложения, то этот фрагмент русского предложения считается переводом фрагмента казахского предложения, и обоим фрагментам (казахскому и русскому) присваивается один и тот же номер.

4. Далее в массиве, сформированном в п. 2, выделяется второй казахский фрагмент, и русские переводные эквиваленты ищутся в русском предложении в направлении слева направо. Если один из этих переводных эквивалентов совпадает с каким-либо фрагментом русского предложения, то этот фрагмент считается переводом второго фрагмента казахского текста, и обоим фрагментам (казахскому и русскому) присваивается один и тот же номер.

5. Процесс установления переводных соответствий между фрагментами казахских и русских предложений должен продолжаться до тех пор, пока не будут исчерпаны все казахские фрагменты массива, сформированного в п. 2. Если при этом для какого-либо казахского фрагмента не удастся обнаружить соответствующего ему фрагмента русского предложения, то этот фрагмент не нумеруется и выполняется переход к следующему за ним фрагменту. Таким образом, может получиться, что не для всех фрагментов казахского предложения будут указаны их переводные эквиваленты на русском языке.

После завершения этапа 3 следует приступить к составлению казахско-русского частотного фразеологического словаря. Единицами этого словаря могут быть сочетания фрагментов казахских и русских предложений, между которыми были установлены переводные соответствия при выполнении этапа 3, а также фрагменты этих предложений, между которыми такие соответствия не были установлены при условии, что они окаймляются фрагментами с одинаковыми номерами. Единицами словаря могут быть и сочетания различных элементов. Далее полученный словарь подвергается проверке, в результате которой исключаются некорректные словарные статьи. Таким образом было получено свыше 34 тыс. фразеологических словарных статей. Фрагмент двуязычного казахско-русского словаря, полученного описанным выше методом, приведен в табл. 5.

Таблица 5

Фрагмент двуязычного казахско-русского словаря

```
**********************************************************
әкімшілік жаза қолдану туралы қаулылардың орындалуы / исполнение постановлений
о наложении административных взысканий
әкімшілік жаза қолдану туралы қаулыны орындау / приведение в исполнение поста-
новления о наложении административного взыскания
әкімшілік жаза қолдану туралы қаулыны орындау жөнінде іс жүргізудің аяқталуы /
окончание производства по исполнению постановления о наложении административ-
ного взыскания
әкімшілік жаза қолдану туралы қаулыны орындауға байланысты әрекеттерге шағым
жасау / обжалование действий в связи с исполнением постановления о наложении
административного взыскания
әкімшілік жаза қолдану туралы қаулыны орындауға байланысты мәселелерді шешу /
разрешение вопросов, связанных с исполнением постановления о наложении адми-
нистративного взыскания
әкімшілік жаза қолдану туралы қаулының міндеттілігі / обязательность постанов-
ления о наложении административного взыскания
әкімшілік жаза қолдану туралы қаулының орындалуын тоқтата тұру / приостановле-
ние исполнения постановления о наложении административного взыскания
```

әкімшілік жаза мерзімдерін есептеу / исчисление сроков административного взыскания
әкімшілік жаза мүліктік залалдың орнын толтыру құралы болып табылмайды / административное взыскание не является средством возмещения имущественного ущерба
әкімшілік жаза сотталушыға қолданылмайды / административное взыскание не может быть наложено на подсудимого
әкімшілік жаза ұғымы және мақсаттары / понятие и цели административного взыскания
әкімшілік жазалардың жекелеген түрлерін орындау тәртібі / порядок исполнения отдельных видов административных взысканий
әкімшілік жазалардың түрлері / виды административных взысканий
әкімшілік жазаның жекелеген түрлерін орындау тәртібі / порядок исполнения отдельных видов административных взысканий
әкімшілік жазаның орындалуынан босату / освобождение от исполнения административного взыскания
әкімшілік жауапкершілік / административная ответственность
әкімшілік жауапқа тартпауға мүмкіндік беретін мән-жайлар / обстоятельства, позволяющие не привлекать к административной ответственности
әкімшілік жауапқа тартылатын адамның түсініктемесі, жәбірленушінің және куәнің жауаптары / объяснения лица, привлекаемого к административной ответственности, показания потерпевшего и свидетеля
әкімшілік жауапқа тартылған адамның кінәсіздігін тану жолымен ақтау / реабилитация путем признания невиновности лица, привлеченного к административной ответственности
әкімшілік жауаптылық / административная ответственность
әкімшілік жауаптылық туралы заңды дұрыс қолданбау / неправильное применение закона об административной ответственности
әкімшілік жауаптылықта болуға тиіс тұлғалар / лица, подлежащие административной ответственности
****

ПЕРЕХОД ИЗ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗА В СИНТЕЗ ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ

На этапе трансфера осуществляется замена понятий исходного текста на понятия выходного языка и преобразование информации о синтаксической структуре исходного текста в информацию, необходимую для синтеза выходного текста. Замена понятий исходного текста на понятия на выходном языке осуществляется с помощью машинного словаря. В результате выполнения этой процедуры выходной текст сначала представляется в виде последовательности слов и словосочетаний, являющихся переводными эквивалентами слов и словосочетаний исходного текста, а затем эта последовательность преобразуется в связный текст. Если в этих группах имеются словарные фразеологические единицы, то запрещается вносить какие-либо изменения в порядок следования входящих в их состав слов, а грамматическая форма слов может изменяться только у опорных слов именных и глагольных словосочетаний и у прилагательных, определяющих опорные слова именных словосочетаний[1, 4].

Традиционно семантико-синтаксический анализ проводится с целью определения синтаксической структуры предложения: установления границ простых предложений, определения главных и второстепенных членов предложения и выявления смысловых связей между ними, построения дерева зависимости предложения и определения для каждого слова однозначной грамматической информации, соответствующей контексту [3]. Разработанный авторами настоящей статьи семантико-синтаксический анализ казахских текстов следует этим принципам. Общая схема процесса семантико-синтаксического анализа казахских текстов приведена на рис. 1.

Синтаксическая структура переведенного текста в значительной степени определяется синтаксической структурой фразеологических словосочетаний, выбранных из словаря. А та часть текста, которая не покрывается словарными фразеологическими словосочетаниями, представляется переводными эквивалентами отдельных слов, которые согласуются друг с другом по правилам грамматики. При этом иногда выполняются локальные (в пределах словосочетаний) и глобальные (в пределах простых предложений) перестановки слов. Для адекватной трансформации словосочетаний на казахском языке в их эквиваленты на русском языке разработано и включено в процедуру казахско-русского трансфера 43 правила преобразования казахского строя языка в русский строй языка.

Результаты машинной реализации процедур семантико-синтаксического анализа и синтеза казахских и русских текстов при переводе одного предложения (в макете используется однобайтовая кодировка символов ANSI, поэтому расширенные кириллические символы кодируются тремя символами с помощью добавления сдвоенных символов ‘ь’ и ‘ъ’) приводятся в табл. 6. Для проверки качества перевода в этой таблице приведен перевод, выполненный человеком-переводчиком.

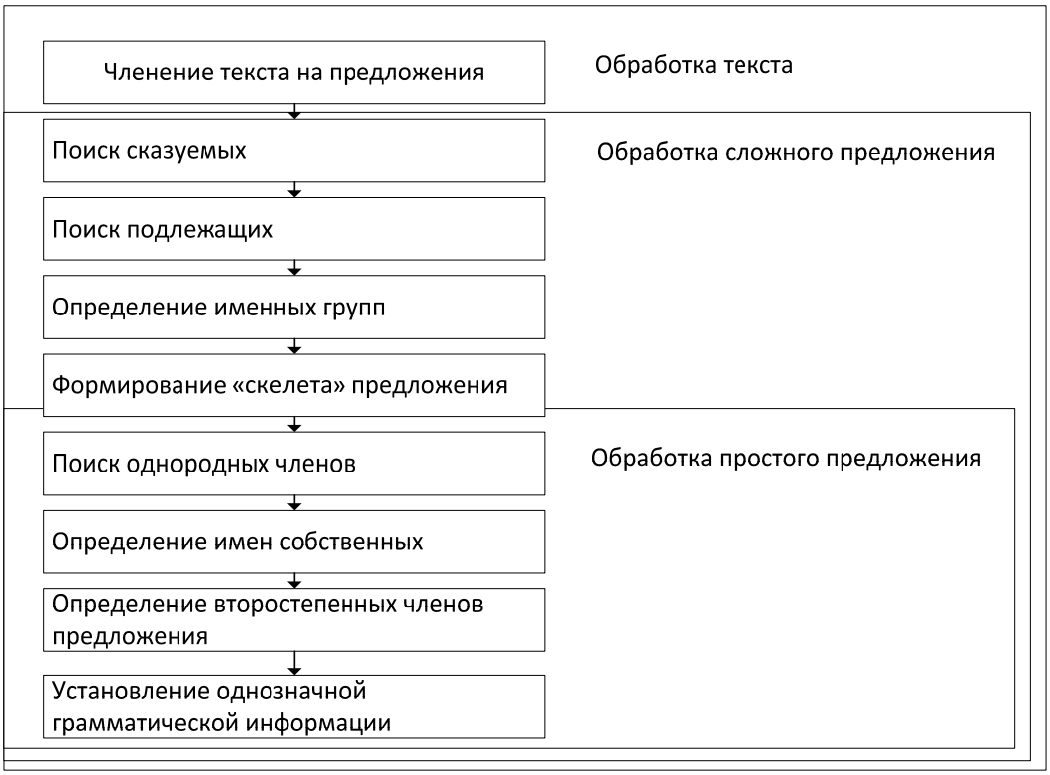

Рис. 1. Схема процесса семантико-синтаксического анализа текстов на казахском языке

Таблица 6

Результаты машинной реализации процедур семантико-синтаксического анализа и синтеза казахских и русских текстов*

Перенумерованный текст исходного предложения
0.Осылайша 1., 2.бизнес 3.деңгейіндегі 4.болашақ 5.салық 6.саясаты 7.ішкі 8.өсімді 9.ынталандыруды 10.және 11.сыртқы 12.нарықтарға 13.отандық 14.экспортты 15., 16.ал 17.азаматтар 18.деңгейінде 19.олардың 20.қорларын 21.,22. Жинақтарын 23.және 24.салымдарын 25.ынталандыруға 26.тиіс 27.
Промежуточный перевод слов и словосочетаний предложения по словарю (Предложение автоматически было разделено на два простых)
Простое предложение №1
#0 таким #1 образом #2 , #3 бизнес #4 на #5 уровне #6 будущее #7 налоговая #8 политика #9 внутренний #10 рост #11 стимулирование #12 и #13 внешние #14 рынки #15 отечественные #16 экспорта #17 , #18 а
Простое предложение №2
#0 граждане #1 встреча #2 на #3 высшем #4 уровне #5 их #6 фондов #7 , #8 накоплений #9 и #10 вкладов #11 стимулировать #12 должен #13 .
Окончательный автоматический перевод казахского предложения
Таким образом, бизнес на уровне будущей налоговой политики внутреннего роста (**должен**) стимулировать внешний рынок отечественного экспорта, а граждане должны стимулировать **(встречи на высшем уровне**) их фонд, накопление и вклад.
Ручной перевод
Таким образом, будущая налоговая политика на бизнес-уровне должна стимулировать внутренний рост и отечественный экспорт на внешние рынки, а на уровне граждан стимулировать их накопления, сбережения и вложения.

* Жирным шрифтом в результатах автоматического перевода обозначены некорректности перевода: отсутствие или неправильный перевод словосочетаний исходного текста.

В табл. 6 показано, что, несмотря на то, что в текущей версии словаря системы МетаФраз [5] имеется значительное количество некорректностей, искажающих как семантико-синтаксическую структуру текстов, так и их переводные соответствия (примером которых служит некорректный перевод казахского слова *деңгейінде (встречи на высшем уровне)* вместо *(на уровне*)), тем не менее, смысл предложения, переведенного автоматически понятен и не противоречит результатам ручного перевода.

СОЗДАНИЕ МАКЕТА СИСТЕМЫ КАЗАХСКО-РУССКОГО ПЕРЕВОДА

На базе разработанных программных средств семантико-синтаксического анализа для казахского языка и модернизированной процедуры трансфера и семантико-синтаксического синтеза русских текстов был создан макет системы казахско-русского перевода (рис. 2).

После создания макета казахско-русской системы перевода была начата его опытная эксплуатация с целью выявления и исправления некорректностей работы системы и пополнения словарей в интерактивном режиме. В течение двухмесячной эксплуатации системы были исправлены критические ошибки и переведены тексты общим объемом 100 Мб. При этом двуязычный словарь системы увеличился до объема 153 тыс. словарных статей.

РЕЖИМЫ МАШИННОГО ПЕРЕВОДА КАЗАХСКИХ ТЕКСТОВ

Автоматический перевод выполняется в многооконном интерфейсе макета казахско-русского перевода. С помощью списков инструментальной панели интерфейса выбираются параметры перевода и подключаются тематический и пользовательский словари. По окончании процесса автоматического перевода его результаты могут быть сохранены в неизменном виде или подвергнуты редактированию в нижнем окне или помощью средств редактора Word. Для удобства пользователя предусмотрена возможность перевода отдельных фрагментов документа.

Процесс интерактивного перевода является продолжением процесса автоматического перевода, в котором пользователю предоставляется возможность только ознакомиться с результатами автоматического перевода всего текста. Интерфейс интерактивного перевода позволяет пользователю корректировать результаты только тех предложений, качество перевода которых его не удовлетворяет. С этой целью ему предоставляется возможность удобного просмотра предложений исходного текста и результатов их перевода. Исходные предложения документа представлены в левом окне интерфейса интерактивного режима перевода, результаты их автоматического перевода – в правом окне.

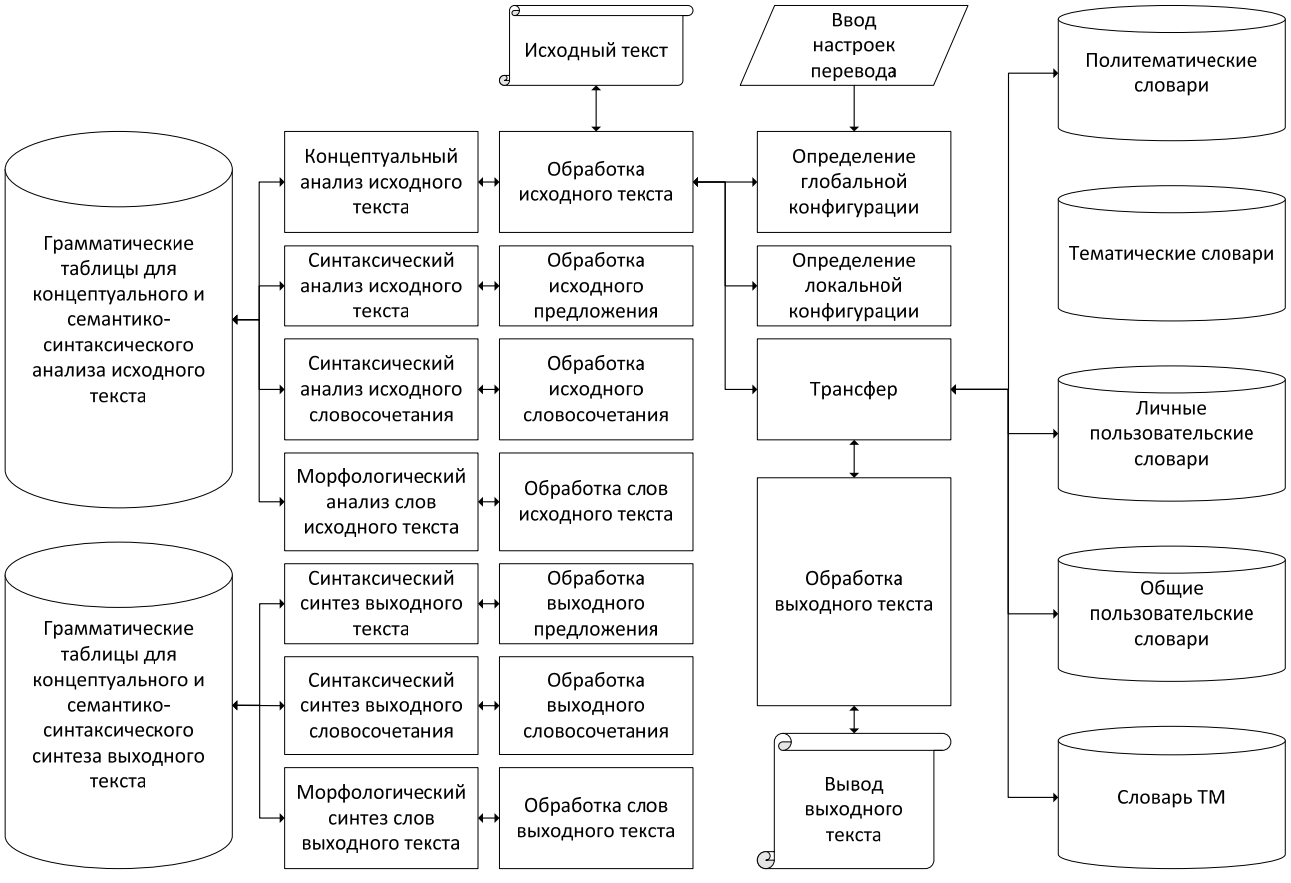

Рис. 2. Общая схема информационно-технологической архитектуры макета системы казахско-русского перевода

Между предложениями исходного текста и их переводом установлена связь. Так, при позиционировании курсора на любом предложении исходного текста в левом окне подсвечивается все это предложение и одновременно подсвечивается перевод этого предложения в правом окне. И, наоборот, при позиционировании курсора на переводе любого предложения подсвечивается исходное предложение в левом окне.

Интерфейс редактирования промежуточных результатов перевода предоставляет пользователю широкий набор инструментальных средств их корректировки. Если пользователь не удовлетворен качеством перевода предложения, то он может воспользоваться некоторыми операциями, позволяющими получить другой, более адекватный вариант перевода. Например, пользователь имеет возможность заменить (операция *выбора*) или изменить (операции *добавления* или *назначения)* переводные эквиваленты слов и словосочетаний, стоящие на первом месте, а также создать новые словосочетания с их переводами из уже имеющихся входных слов и словосочетаний (операция *объединения)*. Кроме того пользователь может, если это необходимо, оставить некоторые слова непереведенными (операция *резервирования)*. Все изменения, вносимые пользователем в промежуточные результаты перевода, немедленно отражаются в нижнем окне интерфейса – результатов перевода предложения и автоматически заносятся в соответствующие машинные словари системы.

С интерфейсными решениями систем фразеологического машинного перевода можно ознакомиться на сайте: http://www.metafraz.ru/

ЗАКЛЮЧЕНИЕ

Для создания казахско-русского направления перевода на базе программно-лингвистической платформы МетаФраз основные усилия были сосредоточены на разработке декларативных и программных средств анализа текстов казахского языка. По результатам этой работы можно сделать следующие выводы.

1. Созданный макет системы машинного перевода с казахского языка на русский позволяет (при ее достаточной адаптации к предметной области) получать информационный (достаточный для понимания основного смыслового содержания) перевод текстов по заданной предметной области.

2. Реализованная модель машинной грамматики казахского языка и модель семантико-синтаксического анализа и синтеза направления казахско-русского машинного перевода полностью адекватна реальным текстам на казахском языке и показала свою состоятельность на тестовых массивах.

3. В соответствии с концепцией ускоренного создания декларативных средств использовались методы формирования словарей и грамматических таблиц без обеспечения полного цикла грамматического и семантического контроля на этапах их создания. При этом предполагалось, что коррекция декларативных средств должна выполняться на этапе технологической адаптации словарной базы к предметной области в процессе опытной эксплуатации.

4. При реализации режима интерактивного перевода казахских текстов были созданы средства автоматизации для обеспечения адаптации всех компонентов словарной базы. Значительная часть некорректностей в декларативных средствах была выявлена и откорректирована в процессе опытной эксплуатации и адаптации макета системы к тематике сайта МИД Казахстана.

СПИСОК ЛИТЕРАТУРЫ

1. Белоногов Г.Г., Калинин Ю.П., Хорошилов Ал-др А., Хорошилов Ал-сей А. Компьютерная лингвистика и перспективные информационные технологии // Научно-техническая информация. Сер. 2. – 2004. – № 8. – С. 30-43; Belonogov G.G., Kalinin Yu.P., Khoroshilov Alexcandr A., Khoroshilov Alexey A. Computer linguistics and promising information technologies // Automatic Documentation and Mathematical Linguistics. – 2004. – Vol. 38, № 4. – P. 28-42.
2. Калинин Ю.П., Хорошилов А.А., Хорошилов А.А. Принципы создания системы мониторинга и анализа мирового потока научно-технической информации // Системы и средства информ. – 2016. – Т. 26, Вып.1. – С. 139–165.
3. Белоногов Г.Г., Хорошилов Ал-др А., Хорошилов Ал-сей А. Автоматизация составления англо-русских двуязычных фразеологических словарей по массивам двуязычных текстов (билингв) // Научно-техническая информация. Сер. 2. – 2010. – № 5. – С. 1-8; Belonogov G.G., Horoshilov A.A., Horoshilov A.A. Automation of the Anglo-Russian Bilingual Phraseological Dictionaries based on Arrays of Bilingual Texts (Bilingual) // Automatic Documentation and Mathematical Linguistics. – 2010. – Vol. 44, № 3. – P. 103-110.
4. Zakharov Victor, Khoroshilov Alexcandr, Khoroshilov Alexey. A Method of Automatic Plagiarism Detection in Multilingual Documents // CEUR Workshop Proceedings Vol-1752. Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016). – 2016. – P. 181-186.
5. Калинин Ю.П., Хорошилов Ал-др. А., Хорошилов Ал-ей. А. Современные технологии автоматизированной обработки текстовой информации // Системы высокой доступности. – 2015. – Т. 11, № 2. – С.19-34.

Материал поступил в редакцию 07.07.17.

Сведения об авторах

ЗАХАРОВ Виктор Николаевич – доктор технических наук, доцент, Ученый секретарь Федерального исследовательского центра «Информация и управле-

ния» Российской академии наук (ФИЦ ИУ РАН), научный руководитель проекта, Москва
e-mail: vzakharov@ipiran.ru

НИКИТИН Юрий Викторович – заместитель генерального директора и системный архитектор лингвистического ПО компании МетаФраз, Москва
e-mail: yuri.v.nikitin@gmail.com

ХОРОШИЛОВ Александр Алексеевич – доктор технических наук, профессор МАИ, ведущий научный сотрудник ФИЦ ИУ РАН, генеральный директор и научный руководитель разработки лингвистического ПО компании МетаФраз
e-mail: khoroshilov@mail.ru

ХОРОШИЛОВ Алексей Александрович – кандидат технических наук, научный сотрудник ФИЦ ИУ РАН, ведущий разработчик лингвистического ПО компании МетаФраз
e-mail: a.a.horoshilov@mail.ru

ИНФОРМАЦИОННОЕ ПИСЬМО И ПРИГЛАШЕНИЕ

МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ К 65-ЛЕТИЮ ВИНИТИ РАН

«ИНФОРМАЦИЯ В СОВРЕМЕННОМ МИРЕ»

Москва, 25-26 октября 2017 г.

подробная информация на сайте: http://www.viniti.ru

Главный организатор:

Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН)

Соорганизаторы:

Российская академия наук
Федеральное агентство научных организаций
Российский фонд фундаментальных исследований
Министерство образования и науки РФ

Проблемно-тематическое направление конференции: современный издательский процесс, интеллектуальная собственность, научные библиотеки, информационное обеспечение научной и инновационной деятельности, информационные технологии для научной и библиотечной отрасли, информационная безопасность, международное сотрудничество и информационный обмен, инфометрия, классификации, стандартизация, образование для отрасли, экономика информации

Основные вопросы, предлагаемые к обсуждению:

● Популяризация научных знаний: Новые модели распространения научной информации
● Редакционно-издательская деятельность в цифровой среде: продукты и сервисы
● Издательские стандарты и технологии
● Перспективы развития книжного дела. Проекты и программы
● Взаимодействие цифровых и печатных ресурсов в научно-технической библиотеке
● Информационно-библиотечное обслуживание: сервисный подход
● Управление данными и навигация в современной научной библиотеке
● Научные библиотечные консорциумы – основные подписчики на научную литературу
● Перспективы развития национальных систем научно-технической информации
● Государственные проекты и программы поддержки информационного обеспечения научно-образовательной деятельности
● Тенденции развития региональных аналитических центров
● Информационное обеспечение экспертной деятельности. Использование информационно-аналитических систем для управления наукой и образованием
● Формальные и неформальные каналы развития современных научных коммуникаций

● Современные агрегаторы научной литературы открытого доступа как источник научной информации
● Машинная обработка данных и аналитические исследования: Приоритеты и сотрудничество
● Использование специальных сервисов компании CrossRef для идентификации научных публикаций
● Роль поисковых систем в современном издательском процессе
● Защита данных от несанкционированного использования. Маркеры безопасности. Политика безопасности открытых систем
● Вопросы достоверности и доверенности при обработке информационного потока
● Межгосударственный обмен научно-технической информацией на евразийском пространстве
● Информационное взаимодействие в рамках СНГ
● Международное партнерство при хранении и обработке больших массивов данных
● Современное состояние систем классификации знаний как инструмента индексирования и поиска данных по перспективным направлениям науки и критическим технологиям
● Современные библиометрические методы определения научных лидеров: Новые математические модели
● Анализ читательской аудитории научной литературы путем вебометрического анализа
● Подготовка специалистов в сфере научно-информационной деятельности
● Мастер-класс по работе с классификационными системами (УДК, ГРНТИ)
● Информация как источник цифрового капитала и фактор социальных изменений
● Информационная деятельность как фактор национальной экономики
● Новейшие бизнес-модели для публикаций открытого и закрытого доступа

На конференции планируются доклады представителей ведущих информационных центров и научно-технических библиотек России, СНГ и дальнего зарубежья.

В рамках юбилейной конференции состоится научно-практический семинар по классификационным системам «Перспективные направления научных исследований и критические технологии в классификационных системах». Предполагается проведение специализированных обучающих мероприятий по УДК индексированию. Запланировано заседание методического совета пользователей ГРНТИ и УДК. Участники конференции получат свидетельства о повышении квалификации.

Материалы конференции будут опубликованы в сборнике Трудов и на CD-ROM, основные – в сборнике **«Научно-техническая информация»**.

Доклады

Принимаются оригинальные работы, имеющие научное и прикладное значение, соответствующие тематическим направлениям конференции и НЕ ОПУБЛИКОВАННЫЕ ГДЕ-ЛИБО РАНЕЕ.

Предлагаемый доклад должен отвечать следующим требованиям:

1. Необходимо указать название доклада, фамилию, имя, отчество (полностью) авторов/соавторов, название организации, город, страну, выделить автора, который будет представлять доклад.

2. Необходимо наличие аннотации, раскрывающей содержание доклада. Размер аннотации - не более 850 знаков (включая пробелы).

3. Доклады принимаются только в электронной форме; тексты – в формате MS Word; схемы, диаграммы, фотографии, сканированные виды экранов и т. п. - в формате JPG. Объем доклада вместе с аннотацией, рисунками, приложениями и т.п. не более 10 страниц формата А4.

4. Доклад необходимо выслать по электронной почте до 11 сентября 2017 г. в адрес оргкомитета: conf@viniti.ru

Доклады, не соответствующие вышеуказанным требованиям,
НЕ РАССМАТРИВАЮТСЯ.

Программный комитет оставляет за собой право определять статус доклада (пленарный доклад, доклад, стендовый доклад), включать принятые доклады в те или иные секции.

Время для выступления: пленарные доклады – 15–20 мин., доклады на отдельных мероприятиях – до 10 мин. Доклады включаются в Труды на основании решения экспертов оргкомитета.

Контакты: 125190, Москва, ул. Усиевича, 20, ВИНИТИ РАН
Телефоны: 8 (499) 152 61 13, 8 (499) 155 42 52, 8 (499) 151 02 61. Факс 8 (499) 943 00 60
Интернет-сайт: http://www.viniti.ru Эл. почта: conf@viniti.ru

База данных (БД) ВИНИТИ РАН

Федеральная база отечественных и зарубежных публикаций по естественным, точным и техническим наукам, генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. Тематическое наполнение соответствует реферативному журналу ВИНИТИ. Для поиска одновременно по всем или нескольким тематическим фрагментам генерируется единая Политематическая БД.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ - http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов за любой период с 1981 г., а также **проблемно-ориентированные выборки** из БД ВИНИТИ по актуальным направлениям научных исследований могут быть предоставлены на договорной основе **в поисковой системе (ИПС) "Сокол"**, работающей под управлением Microsoft Windows и обеспечивающей следующие возможности:

- **Чтение** документов в режиме последовательного просмотра или выборочно по оглавлению за весь период заказанной ретроспективы
- **Поиск** документов по автору, заглавию, источнику, ключевым словам или словосочетаниям, реферату, рубрикам, году издания, стране, языку и т.д. (всего более 20 признаков)
- **Словарь** системы поможет правильно подобрать термины для поиска и выбрать глубину их усечения.
- **Для уточнения поиска** можно дополнительно использовать год издания документа, язык текста документа, рубрики, шифры тематических разделов БД.
- Выполненные **запросы можно сохранять** для их последующего использования и/или редактирования.

125190, г. Москва, ул. Усиевича, 20, БД ВИНИТИ РАН.
Отдел взаимодействия с потребителями – (499) 155-45-25, (499) 152-58-81
E-mail: csbd@viniti.ru, sales@viniti.ru

WWW: http://www.viniti.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

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

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

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

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

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

Подробную информацию Вы можете получить:

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

Коммерческое управление

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

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