DOI: 10.25728/mlsd.2023.0450

О ФОРМАЛИЗАЦИИ КРИТЕРИЕВ ОПТИМИЗАЦИИ MATH-NET.RU: ЦЕЛИ, ВОЗМОЖНОСТИ, СОПРОВОЖДЕНИЕ, ПРИМЕНЕНИЕ

Гончар Д.Р.
ФИЦ ИУ РАН, Москва, Россия
trpl@ya.ru

Аннотация. С целью дальнейшего совершенствования БД Math-Net.Ru предлагаются формальные показатели её полезности и удобства для разных категорий пользователей и сотрудников сопровождения. Обсуждаются некоторые возможные меры по улучшению (оптимизации) этих показателей.

Ключевые слова: информационные системы по поддержке научных исследований, Math-Net.Ru, оптимизация целей, структуры, сопровождения, применения информационной системы.

Введение

В условиях заметного усиления в последние годы ограничений и препятствий, чинимых России в целом и её гражданам в доступе к разного рода передовым технологиям со стороны США, Западной Европы и их союзников, возрастает роль повышения качества и надёжности собственных российских разработок, в том числе и систем информационного обеспечения научной деятельности, таких как Math-Net.Ru (далее также – MN).

В статье отмечается, что на сегодня сопряжение с пользователем в MN позволяет использовать не все возможности поиска по накопленным данным, а накопление ряда иных важных данных (которые собираются в ряде родственных информационных систем, в том числе российской разработки) и вовсе не производится.

Для обоснованности обсуждения возможной оптимизации системы предлагается формализовать ряд критериев качества как самой сегодняшней MN, так и подходов по её ведению и использованию. На этой основе предлагаются ближайшие и перспективные меры и направления возможного совершенствования Math-Net.Ru.

1. Заложенные при создании возможности Math-Net.Ru и степень их использования

Основными сущностями в БД Math-Net.Ru на сегодня являются [1]:

А. ФИО автора.

Б. Название учреждения, где трудится автор.

В. Научный труд (как правило, статья).

Несколько особняком стоит сущность «Журналы», поскольку она является агрегатором подмножества сущности «Научный труд».

1.1. Дополнительные описатели сущности «Автор»

В качестве дополнительных полей к каждой из сущностей имеются следующие основные описания.

А. ФИО автора

- Темы научной работы (текстовое поле)
- Ключевые слова (текстовое поле)
- Научная биография (текстовое поле)
- Учёная степень и учёное звание (из классификатора)
- Год присвоения защиты (поле даты)
- Код ВАК (из классификатора)
- Код УДК
- Ряд полей с гиперссылками (если имеются и известны) на учётные записи автора в иных отечественных и международных научных базах данных, на личную страницу автора, статью о нём в Википедии, список научных трудов автора в отдельном файле (если имеется и был загружен в систему) и т.п.
- Дата рождения (поле даты) и место рождения (текстовое поле)
- Если автора уже нет с нами, то Дата смерти (поле даты)
- Является ли автор зарубежным исследователем.
- Некоторые контактные данные (электронная почта, служебный телефон, факс).

- Ряд служебных полей.

Отметим, что в форме поиска автора при этом на сегодня используются только

- Полное или усечённое ФИО автора (три поля).
- Ключевые слова.
- Организация.

И это, конечно, немало, но запросы, использующие код ВАК, УДК, явное указание какого-то временного отрезка (выхода труда, работы автора в данном учреждении и т.п.) через общедоступное сопряжение с пользователем ныне невозможны.

Не использование содержания кода ВАК в запросах приводит к сниженному вниманию и поддержке разработчиков по ведению соответствующего классификатора в обсуждаемой БД, что, к примеру, заметно по тому, что многие коды общероссийского классификатора в БД отсутствуют. Так, по всем направлениям химических исследований на лето 2023 г. был доступен для выбора только один код. Заметные пропуски в представленности кодов в классификаторе БД имеются и по физико-математическим наукам, а то, что в меню эти коды заметно не упорядочены по возрастанию, делает их заполнение более трудоёмким.

Другим ограничением является возможность указать не более одного кода ВАК для одного автора, хотя нередко даже один и тот же труд (например, диссертация) одновременно имеет несколько кодов ВАК. Вообще говоря, есть и вопрос – где уместнее указывать код ВАК – для автора или для конкретной публикации? Или, возможно, и там, и там?

Понятно, что эти обстоятельства вместе взятые приводят к тому, что у большинства авторов соответствующие поля (ВАК и УДК) на сегодня в MN остаются незаполненными.

1.2. Дополнительные описатели сущности «Учреждение (организация)»

Б. Название учреждения, где трудится / когда-либо трудился автор.

- Собственно название, обычно только полное.
- Юридический адрес, сетевая страница, контактные данные
- Краткая историческая справка о создании и достижениях учреждения.

По поводу представления учреждения в БД Math-Net.Ru заметным недостатком является отсутствие возможности естественным образом указывать соподчинение (иерархичность) подразделений внутри учреждения и отрезок времени с которого и по какой осуществлялась или осуществляется сотрудничество автора с данным учреждением.

Заметим, что в таких российских системах как ИСТИНА МГУ [2] или библиографическая система Института катализа СО РАН [3], представление таких данных предусмотрено. Потребность как-то более точно указать своё подразделение в учреждении (к примеру, факультет в составе крупного университета) очень заметна и в самой Math-Net.Ru. Но на сегодня разные факультеты и иные подразделения одного учреждения представляются в системе на том же уровне, что и учреждение в целом.

К примеру, есть МГУ и на том же уровне – мехмат МГУ, ВМК МГУ, биофак МГУ и т.д. В одном только МГУ ныне около двух десятков факультетов, также свою принадлежность хотят более точно отметить сотрудники НИВЦ МГУ, НИЯФ МГУ и ряда других научных подразделений.

МГУ, конечно, особое учреждение.

Возьмём Московский авиационный институт. При попытке сопоставить автора этому институту и вводу слов «Московский авиационный институт» предлагается 11 разных кодов, обозначающих как разные названия учреждения в разные годы, так и МАТИ им. К.Э. Циолковского (поскольку его воссоединили с МАИ), филиалы МАИ в других городах и отдельные факультеты МАИ.

Разработчики предусмотрели возможность (сотрудникам службы сопровождения) неявно для пользователя указывать синонимичность и вложенность подобных записей, однако при этом теряется ряд сведений и возможностей поиска в системе.

Так, многие институты АН СССР / РАН в разные годы переименовывались, подверглись слиянию, объединению. Поэтому объединение всех сотрудников даже одного учреждения, которые работали в нём при СССР и после 1991 г. заметно, по мнению автора, снижают выразительные возможности поисковых запросов в системе. А если несколько институтов объединили (к примеру, ИСА РАН, ИПИ РАН, ВЦ РАН в 2015 г. в ФИЦ ИУ РАН), то почему нет возможности поиска сотрудника только из ИСА РАН, а идёт поиск только по объединённому списку?

Кроме снижения выразительных возможностей поиска, отсутствие представления вложенности приводит к заметному росту трудоёмкости при заполнении поля «Организация» (приходится выбирать из заметно большего числа похожих учреждений, выдаваемых в меню на одном уровне) и

возникновению положения, когда многие учреждения (даже без указания факультета или иных подразделений) оказываются с несколькими разными кодами.

Автор считает, что многим пользователям было бы удобно

1. Наличие возможности указывать при поиске не просто общее семейство учреждений, имеющих отношение к одному из современных, а более точно. К примеру, чтобы можно было различать ФИАН АН СССР и ФИАН РАН.

2. Переход к явному заданию в MN соподчинения учреждения в целом и его подразделений (факультетов, научных центров и т.п.).

1.3. Дополнительные описатели сущности «Научный труд»

Сущность «Научный труд» (как правило, статья) имеет следующие основные описатели.

- Название труда на русском
- Название труда на английском языке (если есть англоязычный выпуск)
- Аннотация
- Ссылки на соавторов труда (внутри MN).
- Ссылки на название журнала (внутри MN).
- Иные выходные данные статьи (номера страниц и т.п.).
- Ссылка на название учреждения, где трудится авторский коллектив (либо одна на всех, либо по каждому соавтору можно указать несколько учреждений).

Удивительно, но на сей день возможности поиска непосредственно по названию статьи (полному, либо сокращённому) в MN отсутствуют. Можно выйти на статью, зная одного из её соавторов (через поиск авторов), можно, зная название журнала, год и номер выхода статьи открыть её в архиве номеров издания.

Возможно, прямой поиск по названиям статей заметно более трудоёмок и разработчики старались избежать перегрузок системы, но что на первом месте (хотя бы в перспективе и с учётом роста технических возможностей) – удобство пользователя или нагрузка на сервер?

2. Предлагаемые показатели удобства / полезности Math-Net.Ru для пользователей и сотрудников сопровождения системы

Обозначим за $X[i]$ значение i-го показателя удобства / полезности MN для пользователей и сотрудников сопровождения системы и перечислим те из них, которые в настоящее время представляются автору наиболее заметными и существенными.

$X[1]$ **– Классификатор кодов ВАК. Полнота представления наиболее часто встречающихся кодов ВАК**, соответствующих представленных в MN статьям, упомянутым в них диссертациям, основным направлениям научной работы авторов.

Текущие сложности: при том, что коды научной деятельности можно выбирать только из меню классификатора MN, значительная часть кодов (даже по физико-математическим наукам) в последнем не представлена. Из всех кодов из раздела химии представлен только один, отсутствуют множество кодов, соответствующих техническим наукам, также заметно представленным в MN.

$X[2]$ – **Классификатор кодов ВАК. Упорядоченность кодов в** соответствующем **меню** (одного уровня), предлагаемому системой редактирования MN.

Текущие сложности: В соответствующем меню MN на сегодня (2023 г.) коды упорядочены лишь частично, отдельные коды при этом приводятся повторно.

$X[3]$ – **Классификатор кодов ВАК. Иерархичность представления кодов (число уровней вложенности) в** соответствующем **меню** (одного уровня), предлагаемому системой редактирования MN.

Текущие сложности: В соответствующем меню MN на сегодня (2023 г.) имеется один уровень вложенности (самый верхний), в связи с чем (а также неупорядоченностью кодов в меню) имеется неоправданно высокая трудоёмкость поиска соответствующего кода в меню для заполнения в карточке системы.

$X[4]$ **– Классификатор кодов учреждений. Иерархичность представления кодов (число уровней вложенности) в** соответствующем **меню** (одного уровня), предлагаемому системой редактирования MN.

Текущие сложности: В соответствующем меню MN на сегодня (2023 г.) имеется один уровень вложенности (самый верхний), в связи с чем (а также неупорядоченностью кодов в меню) имеется неоправданно высокая трудоёмкость поиска соответствующего кода в меню, повторы содержания в

разных кодах, потеря части данных, существенных для поиска, если для пользователя важна принадлежность автора к определённом подразделению (факультету, научному центру) учреждения, к определённому времени работы этого учреждения (к примеру, до 1917 г., СССР, после 1991, до слияния / разделения с другими учреждениями и т.п.).

X[5] – **Классификатор академических званий. Полнота представления соответствующих званий**, включая звания академиков и членов-корреспондентов республик бывшего СССР (как до, так и после 1991 г.).

Текущие сложности: ряд званий академиков и членов-корреспондентов республик бывшего СССР (как до, так и после 1991 г.), в том числе встречающихся у авторов представленных в MN статей, в соответствующем классификаторе не представлена.

X[6] – **Классификатор академических званий. Число уровней представления**.

Текущие сложности: В соответствующем меню MN на сегодня (2023 г.) имеется один уровень вложенности (самый верхний), в связи с чем замедляется поиск в меню. Заметим, что применение большего числа уровней вложенности позволило бы отражать и академические звания авторов из большего числа зарубежных стран (не только бывшего СССР).

X[7] – **ФИО автора. Полнота заполнения (имени и отчества)**.

Текущие сложности: В карточках авторов MN на сегодня у большой доли авторов (десятки процентов) в полях имени и отчества заполняются только инициалы, что затрудняет поиск по авторам трудов. Наверное, это частично объяснимо, когда речь идёт об отсканированных статьях полувековой давности (и даже более древних). Однако привычка к заполнению только инициалов авторов в ряде изданий стала настолько привычной, что эти поля не заполняются и для авторов, первые публикации которых относятся уже к 2020-м годам нашего века. Автор считает это уже проявлением пониженной общественной ответственности таких издателей перед сообществом авторов и пользователей MN и что этот вопрос надо решать: с какого-то (обозримого) года загружать статьи только с указанием полного ФИО каждого соавтора.

Несколько показателей было бы полезно внести и для оценки качества меню поиска в системе.

Так, для меню поиска по авторам предлагаются следующие показатели:

X[8] – **Доля правильной обработки официально принятых полных и сокращённых названий учреждений** (как дополнительных ограничений при поиске авторов).

Текущие сложности: При попытке указания в форме поиска авторов также названия соответствующего учреждения, многие официальные сокращения (как и полные названия) не обрабатываются системой ожидаемым образом (т.е. пренебрегаются или истолковываются иначе).

К примеру, система распознаёт сокращение МГУ, но уже отказывается понимать «ВЦ АН СССР», «ВЦ РАН», «ИСА РАН». Кстати, «ВМК МГУ» уже не распознаётся (при том что «мехмат МГУ» и «физфак МГУ» системе оказались известны). При попытке указать «биофак МГУ» единственным ответом системы был автор Н.В. Калачёва (карточка № 62197) с биолого-почвенного факультета Казанского ГУ…

При задании более полного имени, к примеру, «Вычислительный центр АН СССР» в качестве ответа выдаётся, к примеру, автор Ю.В. Гончаров (карточка № 53442), все приведённые статьи которого относятся уже к третьему тысячелетию. Но ведь это уже определённое искажение фактов.

X[9] – **Возможности настройки (расширения или сужения состава полей формы поиска автора) из числа полей, имеющихся в MN** (как дополнительных ограничений при поиске авторов).

Текущие сложности: Есть ряд пользователей, которым было бы удобнее использовать в качестве дополнительных при поиске авторов иные поля, чем сейчас заданы в системе.

Например, точную или сокращённую (только год) дату рождения (или ухода) автора, указанную в его карточке специальность ВАК, место рождения, учёную степень и академическое звание. Из международных кодов напрашивается поиск по ORCID, кодам авторов в известных международных БД научных публикаций.

Все эти данные могут быть занесены в систему и уже полностью или частично заполнены для ряда пользователей. Но воспользоваться этим через общий доступ на сегодня не возможно.

Автор считает, что пользователю целесообразно дать возможность настраивать состав полей поиска (на основе имеющихся в MN сведений).

Один из вопросов, которые здесь естественно возникают, это кого мы видим в качестве пользователей системы MN – только и исключительно научных сотрудников в соответствующих областях физики, химии, техники, математики и иных представленных в БД или, начиная с какого-то этапа совершенствования и наполнения системы она может заинтересовать уже и историков, социологов, управленцев науки?

Если так, то какие дополнительные поля следует предусмотреть для повышения успешности и удобства их работы? К примеру, желательно ли ввести отдельное поле института (или даже институтов), которые окончил данный автор. Да, это сейчас можно указать в разделе «Научная биография», но по этому разделу поиск в системе не производится.

$X[10]$ – **Возможности настройки (расширения) состава полей формы поиска автора) при отображении итогов запроса** (как дополнительных сведений при неполных и / или неоднозначных данных по полям выдачи по умолчанию).

Текущие сложности: Состав полей выдачи запросов по поиску авторов ныне включает только поля «Фамилия», «Имя» и «Отчество» автора. При том, что БД включает уже порядка сотен тысяч фамилий и широкой распространённости целого ряда из них, на один и тот же довольно узкий запрос (с полностью указанной фамилией и инициалами) нередко выдаются несколько подходящих сочетаний ФИО авторов. Перебор всех карточек в таких случаях весьма трудоёмок. В ряде других БД подобного назначения (Scopus, РИНЦ [4]) для решения этого вопроса рядом с именем автора отображается его текущая привязка (принадлежность) к учреждению. Почему бы этого не сделать и в MN (можно выводить не одно, а все учреждения, с которыми указана связь данного автора)?

3. Формальная постановка задачи оптимизации удобства / полезности Math-Net.Ru для пользователей и сотрудников сопровождения системы

Будем считать, что для каждого из вышеприведённых показателей $X[1]$ … $X[10]$ предложены определённые разумные интервалы значений, отражающих рост качества (удобства, полезности) системы по данному показателю. Для некоторых показателей такое деление достаточно очевидно (например, число уровней вложенности в том или ином классификаторе), для других их задание уже представляет некую самостоятельную задачу, подробное рассмотрение которой выходит за рамки данной статьи.

Также считаем, что для определённого исторического периода можно с какой-то точностью определить оптимальное значение каждого показателя $XOPT[i]$, (i = 1…10) (быть может, в связи со значениями некоторых связанных с ним показателей), к величине которого желательно стремиться при наличии доступных технических, финансовых, кадровых и иных возможностей для выполнения соответствующих работ.

Тогда обозначая разницу значений по каждому из показателей от оптимального как $Dif[i]$, где $Dif[i] = XOPT[i] - \min(XOPT[i], X[i]), \quad i = 1...10$, а стоимость доведения показателя до оптимального значения как $Cost(Dif[i])$, общие ограничения по затратам как *Budget*, можно записать постановку задачи оптимизации системы Math-Net.Ru для N показателей качества (удобства, полезности) как

$$\begin{cases} \sum_{i=1}^{N} Dif[i] \to \min \\ \sum_{i=1}^{N} Cost(Dif[i]) \le Budget \end{cases}$$

4. Об уровнях сопровождения системы Math-Net.Ru в настоящем и в будущем

В настоящее время в MN можно выделить три крупных уровня сопровождения, наполнения и редактирования системы. Это разработчики и общие хозяева (ответственные) за систему от МИАН, это представители издателей журналов, представленных в MN и это отдельные пользователи.

Первая группа имеет максимум полномочий, но, в силу своей естественной ограниченности по количеству привлечённых сотрудников, наличию у многих из них основной работы собственно в математике, преподавании и т.п., конечно, не успевает и не может успеть выполнить все возникающие задачи на желательном уровне.

Представители второй группы зачастую преследуют близкие цели (размещение очередных выпусков своего издания) и при этом стремятся затратить минимум своих сил, в том числе и за счёт неполного заполнения имён и отчеств авторов, недостаточно внимательного сопоставления вновь загружаемых трудов с имеющимся списком авторов (отсюда не столь уж редко появление нескольких авторских карточек в MN на одного физического автора), а про заполнение дополнительных полей, как учёные степени и звания, ссылки на труды автора в иных родственных по назначению сетях и говорить не приходится.

Третья группа (собственно авторы, которые могут записаться на портале) имеют полномочия править только свою карточку.

Без долгих и во многом очевидных пояснений автор считает, что весьма желательно появление ещё одной группы пользователей – ответственных за ведение авторов по данному учреждению в целом. Ведь нельзя рассчитывать, что каждый автор найдёт время (и воодушевление) записаться на MN, чтобы дозаполнить свои данные, а то и поправить допущенные кем-то при этом неточности. Многих авторов, к тому же, уже нет с нами. Видимо, наиболее естественно искать таких ответственных среди служб учёного секретаря, отдела научно-технической информации. Считаю полезным добавить в этот список и представителей Советов молодых учёных и специалистов (СМУиС) учреждения. У последних энергия молодости, естественное желание познакомиться с историей в лицах учреждения, в котором они оказались, внести посильный вклад в восстановление полных ФИО наиболее интересным им сотрудников своего отдела, отделения и т.п. сочетали бы и увлечение и пользу для науки.

5. Заключение

Среди ряда проектов баз данных научных публикаций, возникших в самой академии наук и притязавших на сколько-нибудь значительный (всероссийский) охват на сегодня осталась одна Math-Net.Ru (разработанная при участии ВЦ РАН система ИСИР РАН [5] вскоре перестала наполняться новыми публикациями, а целый ряд местных проектов [6–8] (и многие другие) обслуживают потребности преимущественно только своего учреждения). Хотелось бы сохранить и приумножить те академические подходы, которые изначально были заложены и пополнялись в MN – открытость для ссылок на сторонние родственные информационные системы, как отечественные, так и международные, стремление дать авторам возможность кратко рассказать о своём пути в науке, иначе говоря, стремление к системе «с человеческим лицом».

MN разработана в МИАН, но уже во многом стала народным научным достоянием страны. Считаю поэтому целесообразным проведение семинаров, конференций по обсуждению и решению накопившихся в MN сложностей, путей их разрешения с более широким кругом пользователей – авторов и читателей Math-Net.Ru. Надеюсь, что это сделает её дальнейшее совершенствование, в котором многие из нас заинтересованы, более быстрым и успешным.

Литература

1. Общероссийский портал Math-Net.Ru. [Электронный ресурс]. URL: https://www.mathnet.ru (дата обращения: 24.08.2023).
2. Научная электронная библиотека eLibrary.Ru – российский информационно-аналитический портал в области науки, технологии, медицины и образования. [Электронный ресурс]. URL: http://elibrary.ru/ (дата обращения: 24.08.2023).
3. ИСТИНА МГУ – Интеллектуальная Система Тематического Исследования НАукометрических данных. [Электронный ресурс]. URL: http://istina.msu.ru (дата обращения: 24.08.2023)
4. Sciact – система мониторинга и учета научной деятельности Институт катализа им. Г.К. Борескова СО РАН. [Электронный ресурс]. URL: http://sciact.catalysis.ru/ru/public (дата обращения: 24.08.2023).
5. Портал ФИАН. БД «Сотрудники». [Электронный ресурс]. URL: http://lebedev.ru/ru/people (дата обращения: 24.08.2023).
6. Институт математики им. С.Л. Соболева СО РАН. Личные страницы сотрудников. [Электронный ресурс]. URL: http://a-server.math.nsc.ru/IM/HP1.asp (дата обращения: 24.08.2023).
7. Вычислительный центр им. А.А. Дородницына ФИЦ «Информатика и управление» РАН. Список сотрудников по отделам. [Электронный ресурс]. URL: http://www.ccas.ru/personal/persor2022.htm (дата обращения: 24.08.2023).
8. *Серебряков В.А.* Электронные библиотеки в вычислительном центре российской академии наук – основные разработки // Электронные библиотеки, 2018. – Т. 21. – № 6. – С. 534–566. [Электронный ресурс]. URL: http://elibrary.ru/item.asp?id=37028487 (дата обращения: 24.08.2023).