СЕЛЬСКОХОЗЯЙСТВЕННАЯ БИОЛОГИЯ, 2015, том 50, № 4, с. 520-524

Новости ЦНСХБ: язык науки

УДК 631.52+636.82]:001.4

МИКРОТЕЗАУРУС ПО ГЕНЕТИКЕ И СЕЛЕКЦИИ СЕЛЬСКОХОЗЯЙСТВЕННЫХ РАСТЕНИЙ И ЖИВОТНЫХ

З.М. ПЕТРАНКОВА, Л.Т. ХАРЧЕНКО, Л.Н. ПИРУМОВА, С.А. ТИМОФЕЕВСКАЯ

Рассмотрены вопросы обработки научной терминологии по генетике и селекции для включения в микротезаурус в Центральной научной сельскохозяйственной библиотеке на основе анализа международных классификаций, тезаурусов зарубежных баз данных и отечественной базы данных «АГРОС». Подчеркивается связь микротезауруса с политематическим тезаурусом по сельскому хозяйству и продовольствию. Представлена технология создания микротезауруса: отбора и обработки лексики, формирования словарной статьи.

Ключевые слова: генетика и селекция сельскохозяйственных растений и животных, лингвистическое обеспечение, информационно-поисковые языки, тезаурусы, сельское хозяйство, Центральная научная сельскохозяйственная библиотека (ЦНСХБ).

Микротезаурусы — один из лингвистических инструментов описания отдельных предметных областей. Они могут быть использованы как основа для согласованной деятельности специалистов, связанных друг с другом одной профессией, научной дисциплиной или областью знаний, при создании, редактировании или индексировании научных текстов. Микротезаурусы создаются по отдельным отраслям сельского хозяйства для удобства индексирования документов (например, в специализированных отраслевых научно-исследовательских учреждениях) и для повышения качества базового политематического Информационно-поискового тезауруса, поскольку позволяют работать со всем объемом терминологии по отдельной предметной области.

Микротезаурус по генетике и селекции сельскохозяйственных растений и животных (МТ) сформирован и ведется на основе политематического контролируемого базового словаря — Информационно-поискового тезауруса по сельскому хозяйству и продовольствию (ИПТ, или тезаурус). В настоящее время в ИПТ содержится более 44 тыс. терминов. Лексический состав тезауруса формировался в течение более 30 лет в процессе индексирования документов и лингвистической обработки терминов различных предметных областей для базы данных «АГРОС». Документографическая реферативная база данных (БД) «АГРОС», генерируемая ЦНСХБ (Центральная научная сельскохозяйственная библиотека), содержит в настоящее время более 1,8 млн записей на отечественные и зарубежные документы (в том числе на статьи из периодических и продолжающихся изданий, тематических сборников) по всем отраслям АПК и смежным областям и наукам.

Технология формирования лексики ИПТ и его производного — МТ включает смысловой анализ текстов при индексировании документов из входного документного потока, поступающего в ЦНСХБ; отбор актуальной терминологии; лингвистическую обработку лексических единиц (ЛЕ); установление между ними смысловых внеконтекстных отношений; редактирование ранее введенных терминов и их отношений в связи с очередным обновлением лексики и другие операции.

В работе по созданию и ведению ИПТ и МТ принимают участие специалисты из разных областей, имеющих отношение к решению научных и прикладных задач АПК, а также лингвисты, специалисты по информационным технологиям и др. Для выработки нормативной лексики привлекаются авторитетные источники (энциклопедии, словари, справочники), проводится анализ документов электронных БД (1-8).

Отметим, что МТ как словарь контролируемой лексики разрабатывается с учетом решения задач информационного поиска в конкретной БД и служит не только сводом нормативной лексики, но и определяет правила индек-

сирования документов предметной области «генетика и селекция сельскохозяйственных растений и животных» (далее генетика и селекция), а также стратегию информационного поиска.

Технология формирования лексического состава разработанного МТ состояла из следующих этапов (9): анализа и отбора лексики из базового ИПТ в электронном формате с помощью набора функций интерфейса и проставления метки принадлежности термина к предметной области «генетика и селекция»; выделения новых терминов по генетике и селекции из массива ключевых слов, использованных при индексировании документов за последние 5 лет, но еще не включенных в ИПТ.

Для выделения новой лексики использовали массив документов БД «АГРОС», сформированный тематическими рубриками отраслевого рубрикатора:

68.03.03.17 Генетика сельскохозяйственных растений
68.03.05.17 Генетика сельскохозяйственных животных. Генетические основы разведения сельскохозяйственных животных
68.35.03 Селекция и семеноводство сельскохозяйственных растений
68.35 Растениеводство (зерновые, технические, овощные, плодовые и другие культуры) (использованы соответствующими подразделами частного растениеводства)

Актуальные термины, как правило, поступают на ввод в ИПТ практически сразу после завершения индексирования документа. Специалист устанавливает смысловые связи термина, предлагаемого для ввода в тезаурус, с другими лексическими единицами (ЛЕ), затем проводятся стандартные процедуры лингвистической обработки, редактирования и ведения ИПТ. Анализ массива документов за несколько лет был предпринят с целью выявления тех терминов, которые индексаторы использовали в статусе ключевого слова, то есть не нормативного термина ИПТ. Критерии отбора ЛЕ — актуальность, частота использования, достоверность (соответствие авторитетным справочникам, базам данных, другим тезаурусам и т.п.).

В области генетики МТ охватывает терминологию общей и прикладной генетики (наследственность, изменчивость, генетический анализ, мутагенез, полиплоидия и др.), основные понятия молекулярной генетики (ДНК, РНК, строение гена, экспрессия генов, транскрипция, генная инженерия, геномика и др.), в селекции, семеноводстве и разведении животных — методы и направления селекции, хозяйственно полезные признаки, сорта и др. Поскольку генетика и селекция тесно связаны с вопросами поддержания новых сортов растений и пород животных, в МТ включена терминология по семеноводству растений и разведению животных (семеноводство, семена, чистопородное разведение, породы животных). Сами же объекты (сельскохозяйственные культуры, полезные растения, сельскохозяйственные животные, лабораторные животные, микроорганизмы) в МТ не включены. Их латинские и русскоязычные наименования со всеми синонимами и связями с другими терминами представлены в настоящее время в базовом ИПТ. Ведется работа по выделению этой лексики в отдельные микротезаурусы.

В основном теле МТ в электронном формате термины представлены в алфавитном порядке латыни и русского языка вместе со всеми элементами лингвистического окружения, которые образуют его словарную статью. В МТ сохранены заданные в базовом ИПТ статус ЛЕ, смысловые отношения между ЛЕ, дефиниции терминов и другие элементы данных. В словарной статье термина могут быть примечание, иноязычный (англоязычный) эквивалент термина, эквивалент термина в международных тезаурусах AGROVOC или CABI, вышестоящие термины (метка «в» с указанием уровня иерархии); нижестоящие термины (метка «н» с указанием уровня иерархии); синонимы (метка «с»); ассоциированные термины (метка «а»). Пример словарной статьи:

семя	Примечание: Для семенного материала исп. семена. Иноязычный эквивалент: seed. Эквивалентный термин в другом тезаурусе: seeds. в1 генеративные органы. н1 семенная кожура. н1 семенные чешуи. н1 семядоли. с семя растений. а семена. а семеноведение

Понятие предметной области может иметь несколько возможных вари-

антов лексического представления. Из них выбирается один термин, называемый дескриптором, который рассматривается как основной способ отображения понятия и используется при индексировании. Дескриптор должен отвечать требованиям общеупотребительности, частоты использования, краткости, терминологической точности.

Для уточнения значения термина, придания ему однозначности используются краткие пометы (реляторы), например: липосомы (органеллы), конкурентоспособность (биол.). Кроме того, значение термина уточняется дефинициями, пояснениями, указанием области применения, которые даются в примечании к термину:

семена	Примечание: Для семени как органа полового размножения растений исп. семя
генетическая трансформация	Примечание: Доставка чужеродных нуклеиновых кислот внутрь интактных клеток. Лежит в основе многих методов генной инженерии

Термины, одинаковые или близкие по значению дескриптору (лексические, условные и другие виды синонимов), а также омонимы, антонимы называются недескрипторами, или аскрипторами. Аскрипторы при индексировании не используются, хотя в авторских текстах и запросах пользователей они могут быть приведены. Между синонимами и дескриптором устанавливаются контролируемые отношения эквивалентности (синонимии, предпочтения). Обогащение МТ синонимами играет важную роль в обеспечении полноты информационного поиска. Пример словарной статьи дескриптора с несколькими синонимами (фрагмент):

плазматическая наследственность	Иноязычный эквивалент: plasmatic inheritance. в1 внеядерная наследственность. с митохондриальная наследственность. с пластидная наследственность. с хлоропластная наследственность. с цитоплазматическая наследственность. а аллоплазматические линии. а материнская наследственность

Следует отметить, что в приведенном примере истинным синонимом является только последний, а первые три — это условные синонимы, поскольку означают понятия, которые входят в объем понятия, отображаемого дескриптором (и могут претендовать в перспективе на роль его нижестоящих дескрипторов). Словарная статья синонима в М имеет вид:

хлоропластная наследственность	Иноязычный эквивалент: chloroplast heredity см. плазматическая наследственность

Синонимичные связи реализуются при автоматизированном редактировании БД, в результате которого происходит замена всех ошибочно или по другим причинам использованных при индексировании аскрипторов на дескриптор. Благодаря этому при поиске в БД пользователь получит все записи по интересующей его тематике, независимо от варианта написания термина авторами документов или в запросе. Например, документы по использованию молекулярных маркеров в селекции будут выданы как по запросу «маркер-вспомогательная селекция» (дескриптор), так и «маркер-вспомогательный отбор» (синоним), а материалы по географическому происхождению каких-либо объектов — как по запросу «географическое происхождение» (дескриптор), так и «место происхождения» (синоним).

Другие виды внеконтекстных смысловых отношений (иерархические и ассоциативные) устанавливаются в МТ только между дескрипторами.

Иерархические отношения (отношения подчинения) устанавливаются между понятиями (терминами), объем одного из которых составляет часть объема другого (род—вид, выше—ниже, целое—часть). Более широкое понятие выражает существенные признаки класса предметов, процессов и т.п., его можно делить на более узкие понятия с учетом общих существенных признаков в качестве основания деления. Классическим примером иерархических отношений, используемых в контролируемых словарях, служат отношения таксономии — древовидной иерархической структуры. Такая структура используется, например, в биологических классификациях. На различных уровнях между основанием и вершиной иерархического дерева находятся термины, каждый из которых

подчиняется только одному термину более высокого ранга (вышестоящему). У него могут быть нижестоящие термины разного уровня иерархии. Например, термины «нуклеиновые кислоты», «нуклеотиды», «нуклеозиды» находятся на одном уровне иерархии и подчинены вышестоящему термину «нуклеиновые соединения», при этом каждый из них имеет подчиняющиеся ему (нижестоящие) термины. Так, иерархическая цепочка термина «нуклеотиды» в МТ имеет вид:

нуклеотиды	в1 нуклеиновые соединения. н1 олигонуклеотиды. н1 пиридиновые нуклеотиды. н1 пиримидиновые нуклеотиды. н1 пуриновые нуклеотиды

Иерархические отношения типа целое—часть, как правило, относятся к физическим объектам, процессам, свойствам, коллекциям и др. В случае МТ это отношения: природные ресурсы—генетические ресурсы; коллекции—генетические коллекции; ядро клетки—хромосомы. Отметим, что некоторые дескрипторы могут иметь 2-3 вышестоящих термина (полииерархия), относящихся к разным иерархическим деревьям или разным ветвям одного дерева:

запасные белки	в1 белки. в1 запасные питательные вещества
горьковская порода овец	в1 мясошерстные овцы. в1 полутонкорунные овцы

Иерархические отношения в значительной степени определяют методику индексирования документов и составления поисковых предписаний (запросов). Они реализуются программными средствами, что обеспечивает приписывание терминам индексирования в поисковом образе документа их вышестоящих терминов. Так, термин «органеллы» автоматически будет приписан всем его нижестоящим терминам, использованным при индексировании:

органеллы	н1 аппарат Гольджи. н1 вакуоли. н1 лизосомы. н1 митохондрии. н1 эндоплазматический ретикулум. с органоиды клетки

Благодаря процедуре приписывания вышестоящих терминов (избыточное индексирование) на запрос «органеллы» (без перечисления всех видов органелл в поисковом предписании) будут получены и документы общего характера, и записи о конкретных видах органелл. Без фиксированных иерархических отношений между терминами выдача записей по запросу будет неполной, в нее войдут только документы общего характера (монографии, обзоры и т.п.), а документы о конкретных видах органелл пользователь не получит.

Ассоциативная связь симметрична и устанавливается между двумя дескрипторами, независимо от их тематической категории или уровня иерархии, но никогда — между терминами, связанными синонимическими или иерархическими отношениями. Ассоциативные отношения между двумя терминами (понятиями, объектами) устанавливаются фактически в случаях, когда при мысленном представлении одного из них возникает другой, связанный с первым в силу какого-либо аспекта его рассмотрения, например, с точки зрения состава, свойств, области применения и др. Например, ассоциативно связаны такие термины: плазматическая наследственность и аллоплазматические линии, материнская наследственность, митохондриальная ДНК, хлоропластная ДНК; цитогенетические маркеры и FISH, GISH, зеленый флуоресцентный белок, полиморфизм хромосом, анеуплоидия, хромосомные аберрации; биохимические маркеры и полиморфизм белков, изоферменты, запасные белки, зеленый флуоресцентный белок.

Основное назначение ассоциативно связанных терминов, приводимых в словарной статье какого-либо понятия, — помочь пользователю при индексировании точнее сориентироваться с выбором нужных дескрипторов.

Остановимся на особенностях представления в МТ некоторых понятий и их индексировании. Так, вводить отдельный термин (словосочетание) «высокорослые сорта» нецелесообразно, поскольку, помимо сортов, могут быть высокорослые клоны, подвои и т.п., кроме того, существуют термины «высокорослость» и «сорта». Поэтому понятие «высокорослые сорта» в МТ рекомендуется при индексировании отображать комбинацией указанных терминов:

высокорослые сорта	Иноязычный эквивалент: tall varieties исп. высокорослость и сорта

(в электронной форме словарной статьи отсылка к использованию комбинации терминов имеет метку =+).

В сформированном МТ содержится более 2 тыс. терминов предметной области «генетика и селекция», из которых около 100 включены в результате анализа массива документов за последние 5 лет. В настоящее время лексика этой предметной области в МТ представлена достаточно полно для отображения при индексировании всех значимых аспектов содержания документов, поступающих на ввод в БД «АГРОС».

Таким образом, микротезаурус по генетике и селекции сельскохозяйственных растений и животных можно рассматривать как средство структурирования и классифицирования понятий и связей, специфичных для данной предметной области; выработки нормативной терминологии и создания согласованного набора терминов для обработки текстов; последовательного использования согласованной терминологии при написании текстов (статьи, монографии, диссертации и др.), их редактировании, выделении ключевых слов, составлении предметных указателей и т.п. Использование МТ способствует повышению качества создаваемых документов, облегчает и ускоряет их обработку для электронных БД, повышает полноту и релевантность информационного поиска.

ЛИТЕРАТУРА

1. AGROVOC. 2014 (http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus).
2. CABI. 2014 (http://www.cabi.org/cabthesaurus/).
3. Multilingual multiscript plant name database. 2014 (http://www.plantnames.unimelb.edu.au /Sorting/List_bot. html).
4. Арефьев В.А., Лисовенко Л.А. Англо-русский толковый словарь генетических терминов. М., 1995.
5. ВИКИПЕДИЯ. Свободная энциклопедия. 2014 (https://ru.wikipedia.org/).
6. Гуляев Г.В., Мальченко В.В. Словарь терминов по генетике, цитологии, селекции, семеноводству и семеноведению. 2-е изд., перераб. и доп. М., 1983.
7. Новый англо-русский биологический словарь. М., 2003.
8. Плантариум. Определитель растений on-line. 2014 (http://www.plantarium.ru/page).
9. Пирумова Л.Н., Харченко Л.Т. Тезаурус по сельскому хозяйству и продовольствию: индексирование документов и поиск информации в БД «АГРОС»: методические материалы. М., 2001.

ФГБНУ Центральная научная сельскохозяйственная библиотека,
107139 Россия, г. Москва, Орликов пер., 3Б,
e-mail: pln@cnshb.ru

Поступила в редакцию
2 марта 2015 года

Sel'skokhozyaistvennaya biologiya [*Agricultural Biology*], 2015, V. 50, № 4, pp. 520-524

MICROTHESAURUS ON GENETICS AND SELECTION OF AGRICULTURAL PLANTS AND ANIMALS

Z.M. Petrankova, L.T. Kharchenko, L.N. Pirumova, S.A. Timofeevskaya

Central Scientific Agricultural Library (CSAL), Federal Agency of Scientific Organizations, 3B Orlikov per., Moscow, 107139 Russia, e-mail pln@cnshb.ru

Received March 2, 2015

Abstract

The genetics and selection terminology processing for inclusion to microthesaurus of the Central Scientific Agricultural Library based on the the analysis of international classifications, foreign databases' thesauruses and local database AGROS are considered. Microthesaurus's connection with polythematic thesaurus on agriculture and food is emphasized. The technology of microthesaurus's creation based on selection and processing of lexicon, formation of lexical entry is presented.

Keywords: genetics and selection of agricultural plants and animals; linguistic support; information retrieval languages; thesauruses; agriculture; Central Scientific Agricultural Library (CSAL).