

ПОПОЛНЕНИЕ КОНТЕНТА ТЕЗАУРУСА КАК СРЕДСТВО ПОВЫШЕНИЯ ЕГО ПОИСКОВЫХ ВОЗМОЖНОСТЕЙ

REPLENISHMENT OF CONTENT OF THE THESAURUS AS A TOOL FOR IMPROVING ITS RETRIEVAL POSSIBILITIES

Пирумова Л.Н., Соколова Ж.В.

Федеральное государственное бюджетное научное учреждение
«Центральная научная сельскохозяйственная библиотека»
E-mail: pln@cnsnb.ru, sjv@cnsnb.ru

Представлены результаты научной работы по актуализации Информационно-поискового тезауруса по сельскому хозяйству и продовольствию, разрабатываемого в ЦНСХБ для обеспечения точного отображения содержания документов в процессе индексирования, унифицированного представления данных, обеспечения описания предметных областей и повышения поисковых возможностей тезауруса. Обозначены основные функции Информационно-поискового тезауруса, являющегося основным информационно-поисковым языком политематической базы данных «АГРОС»: сбор, нормализация и систематизация используемой в научной литературе лексики; индексирование документов и поисковых запросов; обеспечение согласованного, единообразного и формализованного представления информации в базе данных и ее продуктах; обеспечение полноты и точности тематического поиска путем программной реализации иерархических отношений и отношений синонимии; формально-логический контроль терминов индексирования базы данных; функция терминологического справочного пособия. Описана методика работ по актуализации и включению новых терминов в тезаурус, установлению парадигматических отношений между лексическими единицами и формированию иерархических деревьев. В 2019 г. откорректировано (добавлено, изменено, удалено) более 5300 лексических единиц. Введены новые термины по защите растений, зоологии, ботанике, механизации сельского хозяйства, экономике АПК и животноводству. Новая версия информационно-поискового тезауруса содержит 54394 лексических единиц. Из них 34869 дескрипторов и 19525 аскрипторов. Более 21660 лексических единиц являются научными (латинскими) наименованиями организмов (из них 1805 новые). Включено 1617 терминов со статусом синонима, из них 1043 лексические единицы являются научными названиями организмов. Добавлено более 2970 связей между терминами (иерархических, синонимичных, ассоциативных).

Ключевые слова: тезаурусы, информационно-поисковые языки, лингвистическое обеспечение, поиск информации, АПК, базы данных, ЦНСХБ.

Для цитирования: Пирумова Л.Н., Соколова Ж.В. ПОПОЛНЕНИЕ КОНТЕНТА ТЕЗАУРУСА КАК СРЕДСТВО ПОВЫШЕНИЯ ЕГО ПОИСКОВЫХ ВОЗМОЖНОСТЕЙ. Аграрная наука. 2019; (7–8): 77–79.

<https://doi.org/10.32634/0869-8155-2019-330-7-69-72>

Введение

Важнейшей составляющей информационного обеспечения аграрной науки в современных условиях является автоматизированные информационно-поисковые системы (ИПС), базы данных (БД), основой функционирования которых является лингвистическое обеспечение. Поэтому разработка, развитие и совершенствование лингвистического обеспечения является одним из условий успешного и эффективного информационного сопровождения аграрной науки. Тезаурусы являются информационно-поисковым языком (ИПЯ), специально разработанным для поиска в автоматизированных ИПС и БД. Информационно-поисковый тезаурус по

Пирумова Л.Н., Sokolova Z.V.

Federal State Budgetary Scientific Institution "Central Scientific Agricultural Library"
E-mail: pln@cnsnb.ru, sjv@cnsnb.ru

There are given the results of a scientific research on updating the information retrieval thesaurus in agriculture and food developed in the CSAL to provide precise displaying the content of documents during indexing, unified data representation, describing subject areas and improving the retrieval possibilities of the thesaurus. The main functions of the information retrieval thesaurus being the basic information retrieval language of the multitopical database AGROS are specified: capturing, normalizing and systemizing the lexis used in scientific literature; indexing documents and retrieval requests; providing harmonized, unified and formalized representation of information in the database and its products; ensuring the fullness and precision of subject search by software support of the hierarchic relations and synonymy relations; formal and logic control of the indexing terms of the database; function of a terminological reference book. The wording procedures for updating and including new terms into the thesaurus, establishing paradigmatic relations between lexical items and forming the hierarchical trees are described. More than 5300 lexical items have been modified (added, changed, and deleted) in 2019. New terms on plant protection, zoology, botany, agricultural engineering, AIC economy and animal husbandry were introduced. The new version of the information retrieval thesaurus contains 54394 lexical items. Of them 34869 are descriptors and 19525 are ascriptors. More than 21660 lexical items are scientific (Latin) names of organisms (1805 of them are new). 1617 terms having a synonym status were included; of them 1043 lexical items are scientific names of organisms. More than 2970 relations between the terms (hierarchic, synonymic, and associative) were added.

Key words: thesauruses, information retrieval languages, linguistic support, information search, AIC, databases, CSAL.

For citation: Pirmova L.N., Sokolova Z.V. REPLENISHMENT OF CONTENT OF THE THESAURUS AS A TOOL FOR IMPROVING ITS RETRIEVAL POSSIBILITIES. 2019; (7–8): 77–79. (In Russ.)

<https://doi.org/10.32634/0869-8155-2019-330-7-69-72>

сельскому хозяйству и продовольствию (ИПТ) является элементом лингвистического обеспечения автоматизированной библиотечно-информационной системы (АБИС) ЦНСХБ. ИПТ представляет собой постоянно обновляемый контролируемый машинный словарь научных терминов, отобранных с учетом их значимости и частоты встречаемости в документах БД и прошедших специальную лингвистическую экспертизу и обработку. Термины ИПТ (дескрипторы, разрешенные к использованию при индексировании, и аскрипторы — запрещенные к использованию при индексировании) упорядочены по алфавитному принципу с указанием на существующие между ними смысловые связи иерархи-

ческого и неиерархического типа (парадигматические отношения).

В основные функции тезауруса входят: сбор, нормализация и систематизация используемой в научной литературе лексики; индексирование документов и поисковых запросов; обеспечение согласованного, единообразного и формализованного представления информации в БД и ее продуктах; обеспечение полноты и точности тематического поиска путем программной реализации иерархических отношений и отношений синонимии; формально-логический контроль терминов индексирования БД; функция терминологического справочного пособия. Таким образом, ИПТ является средством индексирования, тематического поиска и представления отраслевой научной терминологии.

Как терминологический справочник тезаурус отражает современное состояние науки за счет наличия в нем специальной лексики в формулировках, наиболее часто встречающихся в научных источниках, но при этом не противоречащих классически сложившимся понятиям и формам [1, с. 52–54]. В отличие от обычных терминологических справочников тезаурус представляет научный термин вместе с его выявленными синонимами. Это отличие обеспечивает эффективный поиск в автоматизированных системах с помощью тезауруса. Объединение в одной словарной статье синонимов термина как связанных данных обеспечивает полноту поиска при использовании в поисковом запросе пользователем любого из них. Другими словами, какой бы из синонимов не использовал пользователь в запросе, он получит всю информацию, имеющуюся в системе по данной теме, независимо от того, каким термином данное научное понятие называется в документах или запросе пользователя [2, с. 18–20]. Поэтому стремление к максимальному выявлению синонимии направлено на максимально полное представление отраслевой научной терминологии, что обеспечивает повышение поисковых возможностей ИПТ. Понятийный аппарат тезауруса должен учитывать тенденции развития науки и практики и его лексическая база должна постоянно пополняться, редактироваться, актуализироваться для адекватного отображения содержания документа, что обеспечивает качество индексирования, влияющего на эффективность тематического поиска. От полноты представления отраслевой лексики в тезаурусе зависит напрямую результативность и эффективность поиска. Стремление представить в ИПТ максимально полно отраслевую лексику объясняется необходимостью отыскать в нем конкретный термин и использовать его в поисковом предписании [3]. Наличие термина в тезаурусе обеспечивает релевантный поиск по теме, которую он представляет.

В Федеральном государственном бюджетном научном учреждении «Центральная научная сельскохозяйственная библиотека» (ЦНСХБ) проведена научная работа по пополнению контента ИПТ новой лексикой.

Целью работы являлась актуализация политематического ИПТ для обеспечения точного отображения содержания документов в процессе индексирования, унифицированного представления данных, обеспечения описания предметных областей и повышения поисковых возможностей тезауруса.

Методика

Работа над созданием ИПТ велась в соответствии с ГОСТ 7.25-2001 «Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и

форма представления» [4] и ГОСТ Р 7.0.91–2015 (ИСО 25964–1:2011) «Система стандартов по информации, библиотечному и издательскому делу. Тезаурусы для информационного поиска» [5], а также методики создания ИПТ [6], разработанной в ЦНСХБ. Методикой определены следующие этапы работы по актуализации ИПТ: отбор лексических единиц (ЛЕ); нормализация ЛЕ; построение классификационных схем; редактирование уже вошедших в ИПТ ЛЕ.

Актуализация ИПТ заключалась в редактировании уже существующих в тезаурусе ЛЕ и их словарных статей (выявлении и добавлении новых синонимов, надстройке и расширении иерархических деревьев), в отборе новых научных понятий и терминов, относящихся к сельскому хозяйству и смежным наукам, их семантической и лексической обработке и формировании новых словарных статей, отражающих парадигматические отношения между терминами.

Редактирование заключалось в устранении ошибочных орфографических, синтаксических форм написания ЛЕ, замены статуса термина в словарных статьях, выявлении и пополнении синонимического ряда отдельных понятий, развитии системы парадигматических отношений, обогащении (надстройке и расширении) иерархических деревьев. Отбор новых понятий и терминов производился на основе анализа научных публикаций в российских и зарубежных изданиях с последующим их мониторингом, в ходе которого устанавливалась частотность, т.е. частота упоминания термина в научных публикациях, его развитие и т.д. После достижения определенной частотности термин проходил лексикографическую и семантическую обработку. Для определения статуса и лексической формы проводилась сверка с ГОСТами, терминологическими справочниками. Для гармонизации термина с международной терминологией осуществлялся поиск эквивалента термина в международных тезаурусах по сельскому хозяйству.

Результаты работы

В 2019 г. откорректировано (добавлено, изменено, удалено) более 5300 ЛЕ. Введены новые ЛЕ по предметным областям: ботаника, зоология, механизация сельского хозяйства, экономика АПК, животноводство, а также защита растений.

В терминологической области «Защита растений» значительно пополнена словарная статья *Lepidoptera*: введены новые семейства, роды, виды, дополнены введенные ранее семейства и роды чешуекрылых вредителей сельскохозяйственных культур и лесных пород. Существенно дополнено чрезвычайно важное с экономической точки зрения семейства чешуекрылых *Pyrallidae* (огневки). Введены латинские наименования 62 новых родов и 94 новых видов, относящихся к данному семейству. Значительно расширены и дополнены семейства *Saturniidae* (сатурнии, или павлиноглазки), *Sesiidae* (стекляницы), *Sphingidae* (бражники), *Tineidae* (настоящие моли). Начата работа по пополнению обширного и чрезвычайно важного с хозяйственной и экономической точек зрения семейства *Tortricidae* (листовертки). Введены новые семейства *Schreckensteiniidae* (малинные моли), *Stenomidae*, *Thyatiridae* (пухоспинки, совковидки), *Thyrididae* (окончатые мотыльки), *Tischeriidae* (одноцветные моли-минеры). Начата масштабная работа по детальной разработке и пополнению словарных статей *Rodentia* (грызуны) терминологической области «Зоология». Введены новые семейства *Apodontidae* (аподонтовые), *Batherygidae* (землепо-

вые), *Capromyidae* (ежовые крысы, хутиевые), *Caviidae* (свинковые), *Chinchillidae* (шиншилловые), *Ctenomyidae* (гребнемышинные, ктеномииды, туко-туковые), *Cuniculidae* (паковые), *Dasyproctidae* (агутиевые), *Echimyidae* (колючешиншилловые, колючие шиншиллы, цепкохвостые щетинистые крысы), *Erethizontidae* (американские дикобразы, древеснодикобразовые, иглошерсты), *Gliridae* (соневые, сони), *Heteromyidae* (мешотчатопрыгуновые, мешотчатые крысы, мешотчатые прыгуны), *Hydrochoeridae* (водосвинковые), *Hystriidae* (дикобразовые).

В текущем году осуществлялся ввод научных названий папоротников, многие из которых имеют декоративное значение. Включены новые роды и виды, относящиеся к семействам *Pteridaceae* (птерисовые), *Dryopteridaceae* (щитовниковые, аспидиевые), *Dennstaedtiaceae* (деннштедтиевые), *Aspleniaceae* (костенцовые), *Athyriaceae* (кочедыжниковые), *Cystopteridaceae* (пузырниковые), *Onocleaceae* (оноклеевые), *Polypodiaceae*, *Gleicheniaceae* (глейхениевые), *Nephrolepidaceae*, *Oleandraceae*, *Osmundaceae* (осмундовые, чистоустовые), *Selaginellaceae* (селагинелловые, плауновые), *Isoetaceae* (полушниковые), *Blechnaceae*.

В терминологической области «Механизация сельского хозяйства» добавлено около 20 терминов. Введено около 190 ЛЕ по общей экономике и экономике сельского хозяйства и пищевой промышленности. Терминологическая область «Животноводство» пополнена ЛЕ, относящимися к таким областям знаний, как биохимия, физиология и экология животных, кормление, разведение животных, охотничье хозяйство, рыбное хозяйство, молочная и мясная промышленность. Большая часть введенных ЛЕ относится к разведению сельскохозяйственных и домашних животных.

ЛИТЕРАТУРА

1. Пирумова Л.Н., Соколова Ж.В. Актуализация информационно-поискового тезауруса по сельскому хозяйству и продовольствию // Междунар. с.-х. журн. — 2018. — № 5. — С. 52–54.
2. Костин А.К., Харченко Л.Т., Соколова Ж.В. Отношения синонимии в информационно-поисковых тезаурусах // Библиотечное дело 2008: библиотеки и профессиональное образование в современном обществе: сборник статей по материалам XIII Международной конференции (23–24 апр. 2008 г., Москва). — М.: МГУКИ, 2008. — С. 18–20.
3. Соколова Ж.В. Роль ИПТ в тематическом поиске [Электронный ресурс] // Культура: теория и практика: электрон.

ОБ АВТОРАХ:

Пирумова Лидия Николаевна, канд. пед. наук, заслуженный работник культуры Российской Федерации, заместитель директора Федерального государственного бюджетного научного учреждения «Центральная научная сельскохозяйственная библиотека»

Соколова Жанна Владимировна, старший научный сотрудник Федерального государственного бюджетного научного учреждения «Центральная научная сельскохозяйственная библиотека»

Результатом актуализации явилась новая версия ИПТ, содержащая 54394 ЛЕ (34869 дескрипторов и 19525 аскрипторов). Более 21660 ЛЕ являются научными (латинским) наименованиями организмов (из них 1805 новые). Всего было откорректировано (добавлено, изменено, удалено) более 5300 ЛЕ, из них 2695 латинских терминов. Добавлено 1617 ЛЕ со статусом синонима и более 2970 связей между терминами (иерархических, синонимичных, ассоциативных).

Выводы

В ходе актуализации ИПТ пополнился новыми терминами ЛЕ, расширен синонимичный ряд ряда терминов, что повышает поисковые возможности ИПТ: чем полнее представлена отраслевая терминология, тем легче и эффективнее поиск, чем больше у термина синонимов, тем точнее и полнее поиск (пользователь может использовать любой из них). Таким образом, актуализированная версия ИПТ ЦНСХБ позволяет адекватно раскрывать содержание документов, является эффективным средством индексирования и тематического поиска в БД «АГРОС». Актуализация является средством повышения поисковых возможностей ИПТ. Создание и развитие ИПТ ЦНСХБ соответствует ГОСТам и современному уровню развития тезаурусов. Включение в ИПТ терминологии по всем отраслям АПК позволяет использовать его в качестве общепрофессионального лингвистического средства. Актуализация и пополнение контента ИПТ способствует повышению его поисковых возможностей, поскольку обладая обширным запасом научной лексики, он помогает составлять поисковые предписания максимально конкретно и корректно, сокращать время поиска и получать релевантную информацию. Новая версия ИПТ включена в технологический цикл научной обработки ФГБНУ ЦНСХБ.

науч. журн. — 2018. — Вып. 1 (22). URL: <http://theoryofculture.ru/issues/80/1032>

4. ГОСТ 7.25–2001. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления: издание официальное: введ. 01.07.2002. М.: Изд-во стандартов, 2001. — 18 с.

5. ГОСТ Р 7.0.91–2015. Тезаурусы для информационного поиска: национальный стандарт Российской Федерации: издание официальное: введ. 01.07.2016. — М.: Стандартинформ, 2016. — III, 93 с.

6. Пирумова Л.Н., Харченко Л.Т. Тезаурус по сельскому хозяйству и продовольствию: индексирование документов и поиск информации в БД АГРОС: (методические материалы). — М., 2001. — 69 с.

ABOUT THE AUTHORS:

Pirumova Lidia Nikolaevna, PHD in Pedagogy, Honoured Worker of Culture of the Russian Federation, Deputy Director, Federal State Budgetary Scientific Institution "Central Scientific Agricultural Library"

Sokolova Zhanna Vladimirovna, Higher Senior Officer, Federal State Budgetary Scientific Institution "Central Scientific Agricultural Library"