https://doi.org/10.47612/2791-2841-2022-2-2-19-23
УДК 025.4:63

✉ *Ж. В. Соколова*

Тезаурус как источник нормализованной отраслевой научной лексики

Соколова Жанна Владимировна,
Центральная научная сельскохозяйственная библиотека (Москва, Россия), отдел аналитико-синтетической обработки документов и лингвистического обеспечения, старший научный сотрудник
ORCID ID: *0000-0001-7202-6320*
РИНЦ AuthorID: *200977*
Email: *sjv@cnshb.ru*

Аннотация. Представлены результаты научной работы по актуализации политематического Информационно-поискового тезауруса по сельскому хозяйству и продовольствию (ИПТ), разрабатываемого в Федеральном государственном бюджетном научном учреждении «Центральная научная сельскохозяйственная библиотека» (ФГБНУ ЦНСХБ), для точного отображения содержания документов в процессе индексирования, обеспечения унифицированного представления данных, адекватного описания предметных областей и повышения поисковых возможностей тезауруса. Рассматриваются основные функции ИПТ, представляющего собой сложную терминологическую систему, между элементами которой (лексическими единицами (ЛЕ)) существуют различные виды смысловых отношений. Особо отмечается роль ИПТ как отраслевого справочника научной лексики. В исследовании осуществлено обогащение контента ИПТ новой лексикой по следующим тематическим областям: защита растений, зоология, микробиология, ботаника, лесное хозяйство, растениеводство. Результатом научной работы по актуализации явилась новая версия ИПТ, содержащая 62 682 ЛЕ. Более 27 600 ЛЕ являются научными (латинскими) наименованиями организмов (из них 1 782 новые). Всего было откорректировано (добавлено, изменено, удалено) около 2 700 ЛЕ. Добавлено около 2 800 связей между терминами (иерархических, синонимичных, ассоциативных). Актуализированная версия тезауруса, включающая новую терминологию, дает возможность адекватного описания предметных областей, точного раскрытия содержания документа в процессе его научной обработки и является эффективным средством индексирования и тематического поиска. Создание и развитие ИПТ ФГБНУ ЦНСХБ соответствует современному уровню развития тезаурусов. Объем ИПТ, развитость его словарных статей, представленные в нем смысловые связи терминов, позволяют достаточно полно описывать предметные области, относящиеся к сельскому хозяйству, пищевой промышленности и смежным дисциплинам.

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

Для цитирования: *Соколова, Ж. В. Тезаурус как источник нормализованной отраслевой научной лексики / Ж. В. Соколова // Библ.-информ. дискурс. – 2022. – Т. 2, № 2. – С. 19–23. https://doi.org/10.47612/2791-2841-2022-2-2-19-23*

Статья поступила: 20.07.2022
Статья принята в печать: 26.12.2022
Статья опубликована: 30.12.2022

✉ *Zhanna V. Sokolova*

Thesaurus as a source of normalized branch scientific vocabulary

Zhanna V. Sokolova
Central Scientific Agricultural Library (Moscow, Russia), Department of Analytical and Synthetic Document Processing and Linguistic Support, Senior Researcher
ORCID ID: *0000-0001-7202-6320*
РИНЦ AuthorID: *200977*
Email: *sjv@cnshb.ru*

Abstract. The results of scientific work for updating of the polythematic Information and Search Thesaurus on Agriculture and Foods (IST) developed in the Federal State Budgetary Scientific Institution «Central Scientific Agricultural Library» (FSNSI CSAL) are presented. The work is carried out to accurately represent the content of documents in the process of indexing, providing the unified data presentation, adequately describing subject domains and increasing the searching possibilities of the thesaurus. The main functions of the IST, which is a complex terminological system with different kinds of semantic relationships between its elements (lexical units (LU)), are considered. The role of the IST as a branch reference of scientific vocabulary is particularly noted. During the investigation the IST content has been essentially enriched with new lexis in the following subject domains: plant protection, zoology, microbiology, botany, forest husbandry, agronomy The result of the research work on updating is a new version of the IST containing 62682 LU. More than 27 600 LU are scientific (Latin) names of organisms (of which 1 782 are new). In total, nearly 2 700 LU have been corrected (added, changed, deleted). More than 2 800 links between terms (hierarchic, synonymic and associative) have been added. The updated version of the thesaurus, which includes new terminology, enables adequate description of the subject areas, accurate disclosure of the content of a document during its scientific processing. It is an effective tool for indexing and thematic search. The creation and development of IST of the FSNSI CSAL is in line with the current level of thesaurus development. The scope of the IST, the development of its vocabulary items and the semantic relationships of the terms represented in it make it possible to describe the subject areas related to agriculture, the food industry and related disciplines quite comprehensively.

Keywords: *information search thesauri, information search languages, linguistic support, information search, AIC, databases, CSAL.*

For citation: *Sokolova Z. V. Thesaurus as a source of normalized branch scientific vocabulary. Bibliotechno-informatsionnyi diskurs = Library & Information Discourse, 2022, vol. 2, no. 2, pp. 19–23 (in Russian). https://doi.org/10.47612/2791-2841-2022-2-2-19-23*

The article was received: 20.07.2022
The article was accepted for publication: 26.12.2022
Article published: 30.12.2022

Введение

Для описания какой-либо предметной области всегда используется определенный набор терминов, каждый из которых обозначает или описывает какое-либо понятие или концепцию из данной предметной области. Совокупность терминов, описывающих данную предметную область, с указанием семантических отношений (связей) между ними является тезаурусом [1].

Тезаурус – это искусственный информационно-поисковый язык (ИПЯ), созданный для выражения основного содержания документа с целью последующего поиска в базах данных (БД). В процессе индексирования – представления информации, содержащейся в документе, в свернутом виде – тезаурус помогает индексатору правильно перевести понятия с естественного языка на формализованный язык, тем самым преодолевая такие трудности естественного языка, как синонимия, омонимия, полисемия, неоднозначность выражений [2, с. 113]. Тезаурус представляет собой сложную терминологическую систему, (систему научных понятий с их связями), в которой термины – лексические единицы (ЛЕ) – обогащены связями синонимии и приведены в структуру, не противоречащую системам научных классификаций (наоборот, опираются и базируются на них) и обеспечивающую возможность построения гибких стратегий информационного поиска [3]. Кроме того, тезаурус – это терминологический справочник, отражающий современное состояние науки за счет наличия в нем специальной лексики в формулировках, наиболее часто встречающихся в научных источниках, но при этом не противоречащих сложившимся понятиям и формам [4].

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 [5, с. 27]. Таким образом, тезаурус является средством индексирования, тематического поиска и представления отраслевой нормализованной научной терминологии.

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 [6].

Среди тезаурусов, понимаемых как идеографические словари, в особую группу выделяются информационно-поисковые тезаурусы (ИПТ), появление и развитие которых связано с автоматизацией информационного поиска в середине XX века. ИПТ – это структурированный словарь для контроля лексики, в котором явно и системно определяются основные семантические отношения (эквивалентности, иерархические и ассоциативные) между терминами естественного языка [7]. При этом ИПТ выполняет важнейшую функцию терминологического отраслевого словаря, отображая современное состояние терминологической базы отрасли (АПК).

В Федеральном государственном бюджетном научном учреждении «Центральная научная сельскохозяйственная библиотека» (ЦНСХБ) проведена научная работа по обогащению контента Информационно-поискового тезауруса по сельскому хозяйству и продовольствию новой отраслевой лексикой в соответствии с современным состоянием науки и практики. Целью работы являлась актуализация политематического ИПТ для точного раскрытия содержания документов в процессе индексирования, обеспечения унифицированного представления данных, адекватного описания предметных областей, повышения поисковых возможностей тезауруса и адекватного отображения терминологической базы отрасли.

Обновление контента ИПТ

В процессе актуализации ИПТ по сельскому хозяйству и продовольствию выполнялись следующие работы: пополнение контента ИПТ новой лексикой; установление и развитие иерархических отношений между терминами (построение иерархических деревьев) с учетом внеконтекстных логических связей между отображаемыми ими понятиями; выявление и ввод новых терминов-синонимов, установление отношений синонимии для существующих ЛЕ тезауруса, устранение неоднозначности терминов; установление ассоциативных отношений между терминами в связи с вводом новых ЛЕ, редактирование иерархических связей, замена их ассоциативными в целях рационального расширения поискового образа документа; ввод комментариев к сложным или неоднозначным понятиям; удаление устаревших и ошибочных терминов, их замена, исправление ошибок в написании терминов. ЛЕ тезауруса приписывались так называемые связанные данные, в частности, англоязычные эквиваленты в международных тезаурусах по сельскому хозяйству CABI и AGROVOC, а также

в официальных англоязычных словарях и справочниках. Важнейшей составляющей работы с ИПТ являлась нормализация отобранных для включения в ИПТ терминов, их семантическая и лексическая обработка. Семантическая обработка состояла в выявлении максимально возможно полного синонимического ряда и установлении (выборе) из ряда синонимов одного в качестве дескриптора. Выбор делался на основе анализа публикаций по конкретной теме, выявлении наиболее часто встречаемого в публикациях термина, проверки выявленной формулировки по ГОСТам, энциклопедиям, терминологическим словарям, отечественным и международным базам данных и тезаурусам. Поскольку выбранный термин будет использоваться в качестве дескриптора, т. е. разрешенного к индексированию термина и рекомендуемого к использованию в научных публикациях и научной коммуникации, этот этап работы является важнейшим в процессе актуализации ИПТ. Лексическая обработка включала выбор формулировки с точки зрения написания, падежа, единственного или множественного числа термина в ИПТ.

В исследовании осуществлено обогащение контента ИПТ новой лексикой и иерархическими деревьями по следующим тематическим областям: защита растений, зоология, микробиология, ботаника, лесное хозяйство, растениеводство.

В терминологической области «Защита растений» осуществлялась работа по пополнению чрезвычайно важного с хозяйственной и экономической точек зрения семейства Curculionidae (долгоносики), многие представители которого являются опасными вредителями сельскохозяйственных культур. Введены латинские наименования 84 родов и 101 вида, относящихся к данному семейству. Ранее введенный род Anthonomus пополнен 18 новыми видами. Пополнено новыми родами и видами семейство Tenthredinidae (настоящие пилильщики). Дополнен 22 новыми видами род фитопатогенных грибов Phoma. Расширена словарная статья Phytoplasma (возбудители фитоплазмозов). Всего по защите растений введено около 1300 ЛЕ.

В терминологической области «Зоология» начаты детальная проработка и пополнение словарной статьи Mollusca (моллюски), чрезвычайно обширной группы организмов, насчитывающей около 130 тысяч видов. Введен класс Bivalvia (двустворчатые, или пластинчатожаберные) с относящимися к нему семействами: Anomiidae, Arcidae (арки), Arcticidae, Cardiidae (сердцевидки), Carditida, Chamidae, (хамовые), Corbiculidae, Corbulidae, Donacidae (донациды), Gaimardiidae, Gastrochaenidae, Glycymeridae, Gryphaeidae, Hiatellidae, Laternulidae, Limopsidae, Lucinidae (люциниды), Mactridae, Malleidae, Margaritiferidae (жемчужницевые, или пресноводные жемчужницы), Mesodesmatidae, Myidae, Mytilidae (митилиды), Ostreidae (устрицы), Pharidae (морские черенки), Pholadidae (фоладиды), Pinnidae, Placunidae, Plicatulidae, Plicatulidae, Psammobiidae, Pteriidae, Solenidae, Tellinidae (теллиниды), Teredinidae (терединиды, или корабельные черви) и Unionidae (перловицы, или униониды). Введен также отряд Veneroida и относящиеся к нему семейства Dreissenidae (дрейссениды), Pisidiidae, Semelidae, Sphaeriidae (шаровковые), Trapezidae, Ungulinidae и Veneridae (венериды).

Создана словарная статья Cephalopoda (головоногие). Введен отряд Myopsida (миопсидные кальмары) с семейством Loliginidae (лолигиниды) и отряд Octopoda (осьминоги) с семействами Octopodidae (октоподиды). Добавлен также отряд Oegopsida (настоящие кальмары) с семейством Ommastrephidae (летающие кальмары, или оммастефиды) и отряд Sepiida (сепии, или каракатицы) с семействами Sepiidae (настоящие каракатицы). Всего по зоологии введено около 210 новых ЛЕ.

В терминологической области Микробиология род Rhizobium пополнен 16 видами. Введены новый тип Glomeromycota, класс Glomeromycetes, отряд Glomerales и семейство Glomeraceae.

В терминологической области «Лесное хозяйство» добавлены 25 видов сосны (Pinus), 8 видов ели (Picea), 7 видов лиственницы (Larix) и 16 видов пихты (Abies).

В терминологической области «Растениеводство» введено около 120 терминов, относящиеся к биологии растений, почвоведению, субстратам для защищенного грунта, селекции и сортам, гидропонике, плодоводству и виноградарству, технологиям агрономии.

Результатом научной работы по актуализации явилась новая версия ИПТ, содержащая 62 682 ЛЕ. Более 27 600 ЛЕ являются научными (латинским) наименованиями организмов (из них 1 782 новые). Всего было откорректировано (добавлено, изменено, удалено) около 2 700 ЛЕ. Добавлено около 2 800 связей между терминами (иерархических, синонимичных, ассоциативных).

Отрасли знаний представлены в тезаурусе следующим образом: ветеринария – 16%; пищевая промышленность – 11%; защита растений – 26%, животноводство – 6%; лесное хозяйство – 7%; садоводство и виноградарство – 3%; генетика и селекция – 4%; декоративные культуры – 2%; техническое обеспечение АПК – 2%; экономика сельского хозяйства – 2%; прочие – 21% (рисунок 1).

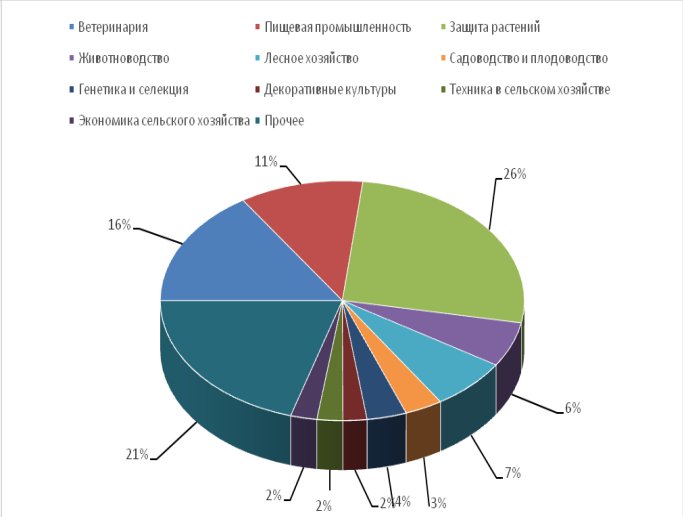

Рисунок 1. – Долевое распределение отраслей знаний в тезаурусе
Figure 1. – Distribution of the fields of knowledge in the thesaurus

Выводы

Таким образом, осуществлено обогащение контента политематического Тезауруса по сельскому хозяйству и продовольствию новой лексикой по следующим тематическим областям: защита растений, зоология, микробиология, ботаника, лесное хозяйство, растениеводство. Актуализированная версия тезауруса общим объемом 62 682 ЛЕ, включающая новую терминологию, дает возможность адекватного описания предметных областей, точного раскрытия содержания документа в процессе его научной обработки и является эффективным средством индексирования и тематического поиска.

ИПТ является средством нормализации отраслевой терминологии. Использование тезауруса в качестве терминологического справочника способствует внедрению стандартизированной, нормализованной, унифицированной лексики. Создание и развитие ИПТ ЦНСХБ соответствует современному уровню развития тезаурусов. Объем ИПТ, развитость его словарных статей, представленные в нем смысловые связи терминов, позволяют достаточно полно описывать предметные области, относящиеся к сельскому хозяйству, пищевой промышленности и смежным дисциплинам.

Список использованных источников

1. Нгуен, М. Описание и использование тезаурусов в информационных системах, подходы и реализация [Электронный ресурс] / М. Нгуен, А. Аджиев // Электрон. б-ки. – 2004. – Т. 7, вып 1. – Режим доступа: https://www.elibrary.ru/download/elibrary_9118514_74383675.pdf. – Дата доступа: 06.07.2022.
2. Онтологии и тезаурусы: модели, инструменты, приложения : учеб. пособие / Б.В. Добров, [и др.]. – М. : Интернет-университет информ. технологий, Бином. Лаб. знаний, 2013. – 176 с.
3. Пирумова, Л.Н. Нормализация и гармонизация отраслевой терминологии в отраслевом тезаурусе: цели и задачи / Л. Н. Пирумова // Плодоводство и ягодоводство России. – 2018. – Т. 52. – С. 181–187.
4. Пирумова, Л.Н. Актуализация информационно-поискового тезауруса по сельскому хозяйству и продовольствию / Л. Н. Пирумова, Ж.В. Соколова // Междунар. с.-х. журн. – 2018. – №5. – С. 52–54.
5. Пирумова, Л.Н. Тезаурус по сельскому хозяйству и продовольствию: индексирование документов и поиск информации в БД АГРОС : метод. материалы / Л. Н. Пирумова, Л.Т. Харченко. – М. : ЦНСХБ Россельхозакадемии, 2001. – 69 с.
6. Пирумова, Л.Н. Пополнение контента тезауруса как средство повышения его поисковых возможностей / Л. Н. Пирумова, Ж.В. Соколова // Аграр. наука. – 2019. – № 7–8. – С. 73–75. https://doi.org/10.32634/0869-8155-2019-330-7-77-79
7. Гендина, Н.И. Информационно-поисковые тезаурусы: структура, назначение и порядок разработки / Н. И. Гендина // Электронный архив НГУ. – Режим доступа: https://nsu.ru/xmlui/handle/nsu/8962. – Дата доступа: 07.07.2022.

References

1. Nguen M., Adzhiev A. Description and use of thesauri in information systems, approaches and implementation. *Elektronnye biblioteki = Russian Digital Libraries Journal*, 2004, vol. 7, iss. 1. Available at: https://www.elibrary.ru/download/elibrary_9118514_74383675.pdf (accessed 06.07.2022) (in Russian).
2. Dobrov B. V., Ivanov V. V., Lukashevich N. V., Solov'ev V. D. *Ontology and thesauri: models, tools, appendices: training manual*. Moscow, Internet University of Information Technologies, Binom. Laboratoriya znanii Publ., 2013. 176 p. (in Russian).
3. Pirumova L. N. Normalization and harmonization of special terminology in the special thesaurus: goals and objective. *Plodovodstvo i yagodovodstvo Rossii = Pomiculture and Small Fruits Culture in Russia*, 2018, vol. 52, pp. 181–187 (in Russian).
4. Pirumova L. N., Sokolova Zh. V. Actualization of information retrieval thesaurus in agriculture and food. *Mezhdunarodnyi sel'skokhozyaistvennyi zhurnal = International Agricultural Journal*, 2018, no. 5, pp. 52–54 (in Russian).
5. Pirumova L. N., Kharchenko L. T. *Thesaurus on Agriculture and Foods: document indexing and information search in DB AGROS*. Moscow, CSAL of the Russian Academy of Agriculture Sciences, 2001. 69 p. (in Russian).
6. Pirumova L. N., Sokolova Z. V. Replenishment of content of the thesaurus as a tool for improving its retrieval possibilities. *Agrarnaya nauka = Agrarian Science*, 2019, no. 7–8, pp. 73–75 (in Russian). https://doi.org/10.32634/0869-8155-2019-330-7-77-79
7. Gendina N. I. Information retrieval thesauri: structure, purpose and development order. *DSpace Repository of Novosibirsk State University*. Available at: https://nsu.ru/xmlui/handle/nsu/8962?locale-attribute=en (accessed 07.07.2022) (in Russian).