

*You say "palta," I say "aguacate" and they say "avocado"*

## Diversity in agricultural terminology of the Americas\*

Lori Finch<sup>1</sup> and Melanie Gardner<sup>2</sup>

### Abstract

A thesaurus is an instrument for controlling words which serves to organize terms and to express relationships among concepts. For decades, information specialists have relied on thesauri to help with the standardization of terminology in information retrieval systems. This article discusses the importance of and need for a joint effort to develop an English/Spanish thesaurus and glossary which reflects the local variations in language found throughout the Americas in the area of agriculture. In 2006, the National Agricultural Library (NAL) of the United States and the Orton Memorial Library (OML) of the Inter-American Institute for Cooperation on Agriculture (IICA) began working together on this effort, and in May 2007 published a bilingual thesaurus. To date, the partners have launched a WIKI, identified an initial workflow and are learning how to work across distances and time zones to create a tool which enhances access to agricultural information across the Americas.

<sup>1</sup> Thesaurus Coordinator, USDA National Agricultural Library, [lori.finch@ars.usda.gov](mailto:lori.finch@ars.usda.gov)

<sup>2</sup> AgNIC Coordinator, USDA National Agricultural Library, [melanie.gardner@ars.usda.gov](mailto:melanie.gardner@ars.usda.gov)

\* We wish to express our gratitude to the comments made by Priscilla Cascante from IICA.

**Key words:** *Thesauri, terminology, agriculture, Latin America, Caribbean.*

## Introduction

Scientists and researchers have struggled with the implications of making their research data and results freely and publicly available through the open access service. The feasibility and availability of information depends on the pervading social, economic, political, contextual and cultural environment. Nevertheless, the amount of scientific information now freely available through open access has substantially increased over the last decade. Alperin, Fischman and Willinsky (2008) note that this trend has been shared by scholars in Latin America and the Caribbean (LAC).

Scientists are making great strides in delivering scientific data and results using the Information and Communication Technologies (ICT) infrastructure. At present, agricultural research offered through open access is now easily disseminated to those trying to solve real problems in the field, such as plant disease diagnosis and control methods.

It is not enough, however, to have more information available via ICT tools if this complicates the search for specific data. With more information available on the Web, it will be more difficult for the researcher, educator, student or scientist to find the data that is needed if he/she does not have the tools required to perform the search. Although open access and the technology to send data to cell phones

*At present, agricultural research offered through open access is now easily disseminated to those trying to solve real problems in the field, such as plant disease diagnosis and control methods.*



facilitate the dissemination of scientific information to those connected to the system, it does not solve the problem of the complexity of language and that of effective retrieval of information.

*The agricultural information derived from the different linguistic variations used by speakers in the Americas also reflects their culture, their need to market and their survival mechanisms.*



## The complexity of language:

The complexity of language, due to the many variations of Spanish spoken in the Americas, creates a multilingual scenario that must be considered in creating agricultural thesauri, as the following example shows:

**Patron:** "Hello, I am looking for articles on fungal diseases of palta."

**Librarian:** "Palta?"

**Patron:** "Persea americana."

**Librarian:** "Oh, yes! Aguacate! Fungal diseases of aguacate. Now I understand."

It is no wonder therefore that the avocado grower in Peru or Chile, who uses the word "palta" for the fruit of the *Persea americana*, does not recognize the term "aguacate," which is used in Central America. It is common for native speakers of Spanish from different regions to not understand the words used outside their own.

From this point of view, the agricultural information derived from the different linguistic variations used by speakers in the Americas also reflects their culture, their need to market and their survival mechanisms.

Considering that the major languages spoken in the Americas are English and Spanish, and that Spanish is the fastest growing language used in the United States in agriculture, there is a need for a standardized bilingual tool that will enable users to use information effectively in adding to and exchanging knowledge in the Americas.

IICA and the NAL have recognized this need and have partnered to develop a bilingual tool that includes as many varieties of the Spanish spoken in the Americas as possible. This resource, entitled *Tesaurus Agrícola*, has been available online since May 2007 and contains over 70,000 terms related to agriculture and ancillary disciplines.

### Facts about NAL Agricultural Thesaurus

- Bilingual Spanish/English
- Emphasizing LAC local terms
- Contains over 70,000 terms
- Scientific and common names of organisms
- Definitions of technical terms available in an additional glossary
- In depth coverage of agriculture and biological concepts
- Special features for indexers such as Enzyme Classification numbers and International Committee on the Taxonomy of Viruses (ICTV) codes
- Collaboration of IICA, BCO, and USDA
- Free file download of XML, PDF, SKOS, MARC formats at the web site, [http://agclass.nal.usda.gov/agt\\_es.shtml](http://agclass.nal.usda.gov/agt_es.shtml)
- Updated annually every January since 2002
- Available on the Internet 24hours a day and seven days a week, with a backup mirror site at Michigan State University


A thesaurus is a tool for controlling words which serves to organize terms and to express relationships among concepts. For decades, information specialists have relied on thesauri to help with the standardization of terminology in information retrieval systems (Gilchrist, Lancaster, Lancaster and Warner).

In addition, the thesaurus acts as a controlled vocabulary in which each term represents one concept. The thesaurus serves as the indexing language for an information retrieval system where it is the set of terms used to express the subject content of items in the information retrieval system. For example:

**Title of article:** “Control of fungal diseases in avocado cultivars grown in the Chanchamayo Valley”

Subject terms from the controlled vocabulary that describe the subject content of the article:

Subject: *Persea americana*  
Subject: avocado  
Subject: fungal diseases  
Subject: cultivar  
Subject: Peru  
Subject: disease control

 *A thesaurus is a tool for controlling words which serves to organize terms and to express relationships among concepts.*

The process of assigning subject terms to items is called indexing. Subject indexing adds value to an information retrieval system so that items are more easily found.

Some of the terminology used in agriculture is a specialized jargon which may not always be clearly understood by speakers outside that field. In the case of some technical jargon, such as biological nomenclature, there are authoritative lists of valid scientific names. This standardization is helpful for communication, such as in the example above. The librarian understood *Persea americana*, which is the scientific name that has been standardized. In the example, the concept of “fungal diseases” can also be expressed several ways, such as:

- *Enfermedades fungosas*
- *Enfermedades micóticas de plantas*
- *Enfermedades por hongos*

To be most effective in finding research papers on this topic, the user would need to use all three phrases in order to find all the relevant information on this topic. Information retrieval systems that use a standard controlled vocabulary such as a thesaurus will make it easier for the searcher to find information since one phrase will be used consistently for the concept of fungal diseases.

The thesaurus offers much more than merely standardization of terms. The thesaurus serves as a way to organize the terms so that “like terms” are together. For example, you will find *alcachofas*, *brócoli*, *coliflor*, *pepinos*, *cebollas*, *tomates* listed together under “*verduras*”.

► Information retrieval systems that use a standard controlled vocabulary such as a thesaurus will make it easier for the searcher to find information.

[See Figure 1] The person searching for information on verduras does not need to recall all verduras as the thesaurus provides a list of them.

Buscar el término	0-9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z	Categorías
<input type="text"/> <input type="button" value="Buscar"/>	<b>verduras</b>  <b>Definición</b> Cualquier parte de una planta que es ingerida comúnmente por los humanos como alimento, pero que no es considerada culinariamente como fruta, nuez, hierba, especia o grano.  <b>English</b> vegetables  <b>Términos Específicos</b> aceitunas aguacate alcachofas alubias verdes apio berenjenas brócoli brotes de bambú brotes de granos calabazas cardo cebollas coles de Bruselas coliflor fruta de pan fruta del pobre gajos de espárragos gombo hinojo hongos comestibles maíz dextrinoso pepinos pimientos puerros ruibarbo tomates tubérculos comestibles vegetales de hoja verde zapallos  <b>Términos Genéricos</b> productos a base de hortalizas  <b>Términos Relacionados</b> cultivos vegetales jardines de verduras jugos de hortalizas	<b>Cambiar de Display</b> Mostrar la Jerarquía . = Términos Específicos : = Términos Genéricos  <b>Busque su término</b> AGRICOLA Artículos AGRICOLA Libros Google Académico
<b>Opciones de la búsqueda</b> Lengua español / Spanish Método de búsqueda Términos que incluyen.. Número de términos desplegados 100		

Figure 1. Excerpt from Tesauro Agrícola demonstrating that like terms are grouped into hierarchies.

The thesaurus also brings together terms which are synonyms. Palta and aguacate are examples of regional equivalents for the same concept. In a controlled vocabulary, one term is the preferred term; that is, the one that is chosen to be assigned for subject description of items in an information system. If we say, “palta USE aguacate,” we are instructing those describing items to use the term aguacate.

Another type of synonymy that is handled in the thesaurus is spelling variants, as shown by the example *turfgrasses* or *turf grasses* [See Figure 2]. These are common in the English language, and are particularly prevalent in the differences between American and British English, such as *oestrogens* and *estrogens*. In the thesaurus, one is chosen as the preferred term and the other is designated as a cross reference.



**Figure 2. Excerpt from Tesouro Agrícola demonstrating spelling variants in English.**

The thesaurus is not a static document, but rather a dynamic resource which must keep in step with agricultural discoveries and technical progress. The NAL and IICA, through the staff at the

Orton Memorial Library (OML), are collaboratively expanding the vocabulary to better accommodate the needs of Latin America. However, there is an urgent need to have experts from across

► *The thesaurus is not a static document, but rather a dynamic resource which must keep in step with agricultural discoveries and technical progress.*

LAC to contribute their regional dialect to this resource so that this knowledge can be shared.

It is essential that all countries contribute so that their research can be found by others and reused to the benefit of all.

IICA and the NAL recognize the need to work together to ensure that this vocabulary tool will be of use in more effectively indexing agricultural literature and provide for improved retrieval of data on tropical and temperate agriculture for agriculturists throughout the Americas.

Adding information to the thesaurus is necessary, but also the review of existing information by language and agricultural subject experts is needed. Since language contains homographs, it is easy for a translator to misunderstand the concept at hand and provide an incorrect translation. For example, seeing the English term “bits” does not convey the meaning behind the term. The translator would need to consult the hierarchy and other notes associated with the term to provide the correct translation.

The original translation of the Tesauro Agrícola was done by a group from Chile. The translator provides translations that are common to that region, but may not represent all the regional varieties in LAC. Review by experts from different countries of LAC is needed so that there is equal and complete representation of regional dialects.



◀ *Maintaining a thesaurus involves subject experts in agriculture, but also specialists in lexicography, and those involved in his process also need to learn about the principles of thesaurus construction.*

In 2008, IICA and NAL began using a collaborative WIKI in order to facilitate thesaurus development and maintenance. The WIKI is a new Web 2.0 tool that makes the collaborative development of knowledge on topics of common interest possible. It is a space for “experts” to provide feedback on agricultural terminology, with a view to reviewing and updating the agricultural thesaurus produced by the NAL. The

WIKI serves as a “white board” where participants can propose new terms for the thesaurus, correct errors in the thesaurus, suggest definitions for terms, and post translations. The proposals discussed over the WIKI will be included in the 2010 edition of the thesaurus. During its first year of use, the WIKI has been useful for establishing workflows, maintaining the thesaurus and allowing the participation of other agricultural subject experts in LAC.

Maintaining a thesaurus involves subject experts in agriculture, but also specialists in lexicography, and those involved in his process also need to learn about the principles of thesaurus construction. General guidelines are established by International Organization for Standardization (ISO) for such construction.

Specific guidelines and principles for construction of the *Tesaurus Agrícola* are established by the NAL with the aid of their IICA partners. Principles and rules need to be established so that there is consistency throughout the thesaurus. For example, rules are established for the treatment of abbreviations, acronyms, symbols, punctuation, capitalization, scientific names, common names, geographic terms, term form and disambiguation of homographs.

The process of the selection of terms is critical and must also follow an established norm. Terms must represent concepts that are well accepted in the discipline. One can find evidence of this acceptance by searching existing literature, such as Agri2000, AGRICOLA or Google Scholar, and determine its frequency of use.

Another method of finding suitable terms is to examine search query logs of users, called user warrant. A term can be justified to be added to the thesaurus if it is frequently searched by users of an information system. Indexers, or those who apply the controlled vocabulary, are excellent sources of new terms. New terms and concepts that are being generated in a discipline are seen by the indexers as they apply the controlled vocabulary. It is advisable to consult with indexers, or analyze the uncontrolled subject terms assigned by indexers, to find candidate terms for inclusion in the thesaurus. The importance of knowing the rules established for constructing thesauri whose structure and content are consistent cannot be underestimated.

## Challenges:

The NAL and IICA will rely on the expertise and leadership of its staff to expand the project to include a wide variety of experts from the LAC. Currently, the OML and the NAL are working to find the best ways to be efficient in their processing of proposals for new terms and changes to the vocabulary. In addition, they are evaluating technologies such as the WIKI to determine their usefulness as an appropriate technology for collaboration on such a geographically distributed project. It is possible that there are other technologies that are needed, especially for the training of participants on the mechanics of thesaurus maintenance. The NAL and the OML are engaged in analyzing which areas in the thesaurus need further development and are identifying

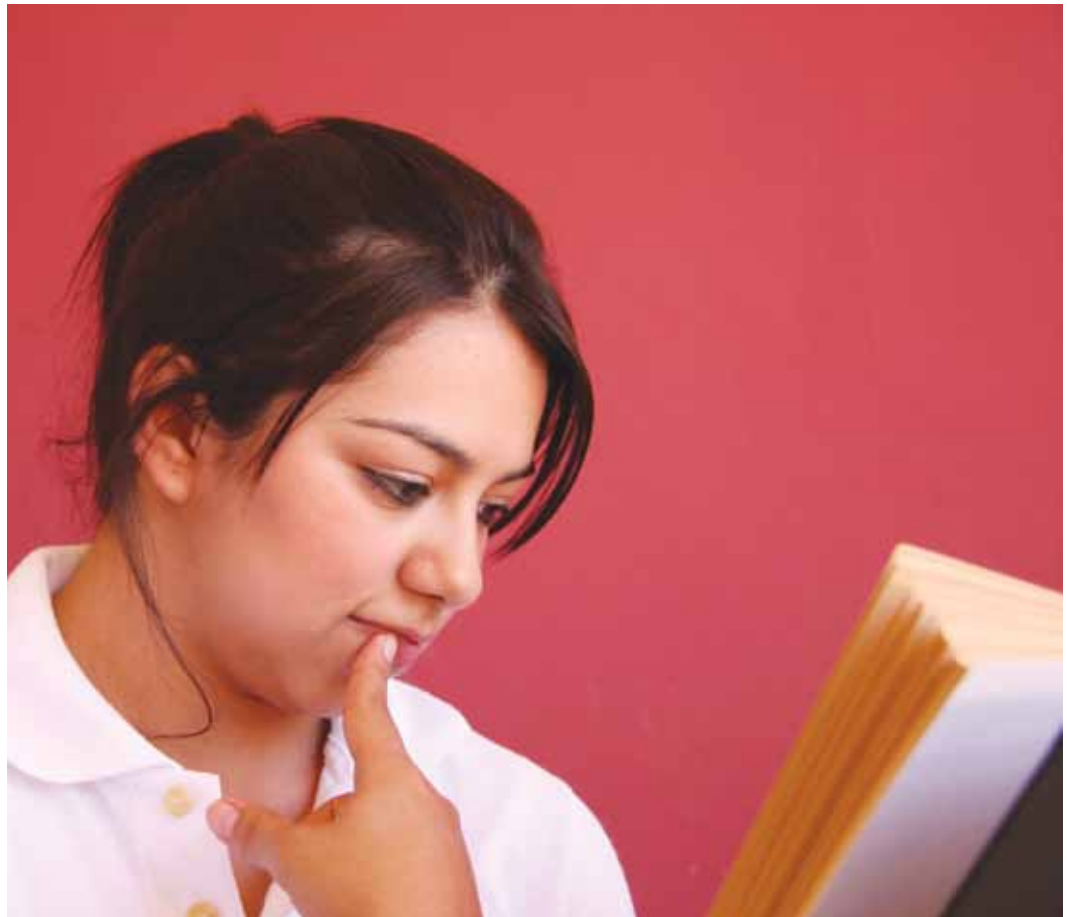


and seeking experts needed for a favorable outcome.

The challenge and opportunity for the future of the expansion and development of the *Tesaurus Agrícola* is imminent. The human resources and intellect needed to do this work is not found in one organization or in one particular country of LAC. Collectively, there is an abundance of knowledge, talent and specialized agricultural expertise in the Americas. It is hoped that the interest in the success of this project, which can

benefit many, will be as strong as the commitment shown by some individuals and organizations in LAC.

We hope that this discourse has inspired some to truly appreciate that the complexity of language is real and deserves our time and attention. It challenges us to cultivate it and form it into a tool that will serve the agricultural information systems of the Americas. This tool will contribute to understanding whether you are the agriculturalist growing avocado in Peru, Guatemala or the United States.



## Literature consulted

- Ager, S. n.d. Latin American Spanish or Spanish for Latin America. Omniglot. Available at [http://www.omniglot.com/language/articles/latin\\_american\\_spanish.htm](http://www.omniglot.com/language/articles/latin_american_spanish.htm)
- Alperin, JP; Fischman, GE; Willinsky J. 2008. Open Access and Scholarly Publishing in Latin America: Ten flavours and a few reflections. *Liinc em Revista* 4(2): 172-185. Available at <http://www.ibict.br/liinc>.
- Chan, LM; Richmond, PA; Svenonius, E. 1985. *Theory of Subject Analysis: A Sourcebook*. Littleton, Colorado, US, Libraries Unlimited, Inc. 415 pp.
- Chamis Yanosko, A. 1991. *Vocabulary Control and Search Strategies in Online Searching, New Directions in Information Management* no. 27:121. New York, US, Greenwood Press.
- Gilchrist, A. 1971. *The Thesaurus in Retrieval*. Londres, Aslib, 183 pp.
- International Standards Organization. *Documentation - Guidelines for the establishment and development of monolingual thesauri*. ISO 2788-1986. 32 pp.
- \_\_\_\_\_. *Documentation – Guidelines for the establishment and development of multilingual thesauri*. 61 pp.
- Lancaster, FW. 1986. *Vocabulary Control for Information Retrieval*. Arlington, Virginia, US, Information Resources Press. 270 pp.
- \_\_\_\_\_; Warner AJ. 1993. *Information Retrieval Today*. Information Resources Press, 341 pp.
- \_\_\_\_\_. 1998. *Indexing and Abstracting in Theory and Practice*. Champaign, Illinois, US, University of Illinois. 412 pp.
- Onsrud, H. 2004. *Overview of Open-Access and Public-Commons Initiatives in the United States. Proceedings of an International Symposium on Open Access and the Public Domain in Digital Data and Information for Science hosted by the Board on International Scientific Organizations*.
- Suber, P. 2007. *Open Access Overview: Focusing on open access to peer-reviewed research articles and their preprints*. Available at <http://www.earlham.edu/~peters/fos/overview.htm>
- Wellisch, HH. 1991. *Indexing from A to Z*. Bronx, New York: HW Wilson Company. 461 pp.

# Résumé / Resumo / Resumen

## ► « Pour désigner l'avocat, vous dites 'palta', je dis 'aguacate' et d'autres, 'avocado' » : La diversité dans la terminologie agricole des Amériques

Un thésaurus est un répertoire structuré qui sert à organiser des termes et à établir des relations entre des notions. Depuis des décennies, les spécialistes de l'information s'appuient sur des *thesauri* pour normaliser la terminologie des systèmes d'extraction d'information. Dans le présent article, nous examinons les fondements et la nécessité du nouveau partenariat qui s'est donné pour tâche d'établir un thésaurus et un glossaire anglais/espagnol rendant compte des variantes locales utilisées dans les pays d'Amérique latine et des Caraïbes. La *National Agricultural Library* (NAL), l'Institut interaméricain de coopération pour l'agriculture (IICA) et la Bibliothèque commémorative Orton (BCO) ont commencé à travailler de concert dans ce but en 2006. Un thésaurus bilingue a été publié en mai 2007. À ce jour, les partenaires ont lancé un WIKI et défini un ordonnancement des tâches, et ils apprennent à travailler à travers les distances et les fuseaux horaires pour créer un outil qui facilite l'accès à l'information agricole dans toutes les Amériques.

## ► Você diz palta, eu digo abacate e eles dizem avocado: Diversidade na terminologia agrícola das Américas

Tesouro é um vocabulário estruturado que serve para organizar termos e expressar relações entre conceitos. Especialistas em informação há décadas têm recorrido aos tesouros para ajudá-los na padronização da terminologia nos sistemas de recuperação da informação. Este artigo discute as bases e a necessidade desse novo esforço conjunto para desenvolver um tesouro e um glossário inglês/espanhol que reflita as variações locais representadas em toda a América Latina e o Caribe. A *National Agricultural Library* (NAL) dos Estados Unidos, o Instituto Interamericano de Cooperação para a Agricultura (IICA) e a Biblioteca Conmemorativa Orton (BCO) iniciaram em 2006 uma parceria nesse sentido, havendo sido lançado um tesouro bilíngue em maio de 2007. Recentemente, esses parceiros lançaram um WIKI, identificaram um programa de trabalho inicial e estão aprendendo a lidar com as distâncias e os fusos horários para criar uma ferramenta que intensifique o acesso à informação em agricultura nas Américas.

## ► "Usted dice "palta", yo digo "aguacate" y ellos dicen "avocado": Diversidad en la terminología agrícola de las Américas

Un tesouro es un instrumento de control de palabras que se utiliza para organizar términos y expresar relaciones entre conceptos. Durante décadas, especialistas de la información han dependido de los tesouros para contribuir a estandarizar la terminología en sistemas de recuperación de datos. En este artículo se discute la importancia y la necesidad de realizar un esfuerzo conjunto para elaborar un tesouro y glosario inglés/español que refleje las variaciones locales del lenguaje utilizadas en los países de ALcen materia agrícola. En el 2006, la Biblioteca Agrícola Nacional de los Estados Unidos (NAL) y la Biblioteca Conmemorativa Orton (BCO) del IICA empezaron a trabajar conjuntamente en ese esfuerzo y, en mayo del 2007, publicaron un tesouro bilingüe. Hasta la fecha, dichas instituciones han lanzado un WIKI, han identificado un flujo de trabajo inicial y están aprendiendo a trabajar a través de las distancias y los husos horarios para crear una herramienta que aumente el acceso a la información agrícola a lo largo y ancho de las Américas.