Annals of Library and Information Studies
Vol. 58, Sept 2010, pp. 278-281

Agrotags: a contribution towards improved digital information management in agricultural research

T.V. Prabhakar[1], Lavanya Kiran Neelam[2] and V Balaji[2]
[1]Department of Computer Science, Indian Institute of Technology Kanpur, India 208016
[2] International Crops Research Institute for the Semi-Arid Tropics, Patancheru, India 502324
Author for correspondence: Prof. T V Prabhakar, tvp@iitk.ac.in

Agricultural research generates highly locale-specific as well as abstract or location-neutral information. Because of this diversity and the need to build prescriptive practices out of such information, agricultural research information management on the web requires semantic enablement. Keyword assignment is an important step towards achieving this. In this paper we describe a taxonomy called Agrotags which is designed for tagging agriculture research documents. Agrotags is a broad subset of Agrovoc, the International thesaurus of the UN Food and Agriculture Organization (FAO) and is much smaller - about 2,100 as against 40,000 terms. Agrotags is manually created by carefully examining each of the Agrovoc terms for their utility in tagging. This selected subset is further refined and validated by looking at the manually assigned keywords from the FAO AGRIS databases. This is used in a new platform called OpenAgri that can be used to build institutional or network-based repositories of agricultural research documents.

Introduction

Eugene Garfield pioneered efforts on mapping of knowledge domains using information available in the *Web of Science*. His *HistCite* enables vocabulary analysis which can help a searcher identify the most significant work on a topic and trace its year-by-year development[1]. On an independent track, we have been developing information management tools in the domain of agricultural extension and research, and we appreciate the insights that Garfield's work in efforts such as the *HistCite* can offer.

The near absence of appropriate information on distinct practices of agriculture/farming in the world of Web 2.0 has already been pointed out in many instances[2]. Efforts like the Agropedia[3] have initiated strategies and created pathways to address this web information paucity by bringing quality extension materials onto the web. These are stored in a manner that facilitates reuse in various contexts and across diverse delivery media.

Information management of agricultural research documents is yet to achieve the level of coherence and sophistication of disciplines such as particle physics. There is no paucity of research reports, published research papers and other documentation related to agricultural research on the web unlike the situation with extension material. Many reputed publishers host many of these articles in their repository. A small number of these repositories use various tagging methods to label documents to facilitate retrieval; while others prefer to let search engines index their repository. The inherent drawbacks of both these approaches lie in the lack of ability to retrieve the documents from the tags. This greatly limits the participation and availability of the document across a semantic network[4].

The need for an ontology-backed tagging methodology was strongly felt partly because of the adoption of agricultural sciences, which span highly locale-specific work as well as abstract information. The combination of advanced metadata tagging, and cross-linking facilitated by controlled ontologies would give rise to a wealth of semantically-linked relevant documents. Many international agricultural thesauri exist like the Agrovoc (a multilingual agricultural thesaurus, http://aims.fao.org/website/Agrovoc-Thesaurus/sub), the CAB thesaurus (http://www.cabi.org/cabthesaurus/), NAL (http://agclass.nal.usda.gov/dne/search.shtml) and MeSH (Medical Subject Headings, http://www.nlm.nih.gov/mesh/MBrowser.html).

Agrovoc, existing since 1976 as a thesaurus, and its morphing into a full-fledged agricultural ontology in the last decade, was seen as a natural choice for the creation of Agrotags. The advantages offered by a semantically-

tagged knowledge repository for agriculture was already ascertained by efforts such as the Agropedia, where the Agrovoc has provided the glue for the semantic inference.

The International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) has long been involved with the Agrovoc enrichment together with the FAO (Food and Agriculture Organization of the United Nations). Indian Institute of Technology Kanpur (IIT Kanpur) has been maintaining the Hindi version of Agrovoc. ICRISAT has led the revision and refinement of the Agrovoc thesaurus which forms the basis of the Agricultural Ontology Service (AOS). Agrotags was envisaged as a collection of terms that would be used to tag digital information objects (DIOs) in the agriculture realm. The main aim is to normalize tagging process in order to make for more efficient and simpler searching and to provide the most efficient resources to the user.

Agrotags's pedigree has thus been Agrovoc. The ongoing efforts to enrich Agrovoc to ontology are widely known[5]. Agrovoc is also working on mapping onto leading thesauri such as NAL, CABI and MeSH (see above). This provides documents tagged with Agrotags rich interconnection with documents tagged with terms from the other thesauri. The inherent power of Agrovoc to convert a term into 19 languages provides an added advantage. Applications built using Agrotags as an assisting-knowledge layer would have therefore greater reach.

Ontogenesis of Agrotags

The development of Agrotags was started by analyzing various tagging options available for research documents especially in the agriculture realm. The inherent drawback was realized as documents tagged in other languages were not 'retrievable' using the tags supplied. An immediate solution lay in the use of terms from Agrovoc, which contains (as of July 2010) almost 40,000 terms in the English language alone - a huge candidate set for generation of tags. The subject matter experts from ICRISAT and IIT Kanpur decided that a collection of hand-picked terms would go into the creation of a collection of terms for tagging agricultural research documents.

Initially, the top term creation was based on popular thesauri like NAL, CAB and MeSH, but later it was decided to create a hierarchy rooted in the concepts from the knowledge models used in agropedia. After the top terms were finalized, the team set about creating the hierarchy taking care to retain the intended purpose of Agrotags, namely to semantically tag research documents. Terms were also sourced outside the Agrovoc to arrive at a comprehensive collection of tags.

Navigating through the 25 top terms of Agrovoc, the team selected terms that were useful for tagging. For example, outbreeding, cultivar selection, mass selection, control methods, etc. are narrower terms of Agrovoc top term methods with different depth levels. However, "outbreeding" and "mass selection" are associated with "crop improvement", "cultivar selection" with "plant production" and "control methods" with "plant protection" which are the top terms of Agrotags (as mentioned). Terms outside Agrovoc were also included in Agrotags and unique codes were assigned to these terms. In the first phase of Agrotags, almost 15 top terms were created. Plant production, plant protection, crop improvement, etc. formed some of the top-level terms. For each of these top-terms, a hierarchy was created from the remaining 1,500 terms which can be viewed graphically at http://agropedia.iitk.ac.in/agro_tag/agro_tree.html. Currently Agrotags are available in English, Hindi and French. Telugu and Kannada versions are in progress.

Criteria of selection

Only descriptors and more popular terms were selected to create Agrotags from the Agrovoc. The non-descriptors (synonym of the descriptor that refers the user to it but may not be used for indexing and searching), scientific/taxonomic names, fishery-related terms and geographical terms were not included in the selection process .This can be elaborated taking into account some simple examples:

1. 'Groundnuts' is a term in Agrovoc (termcode-11368) and has a non-descriptor 'peanuts'. 'Groundnuts' is a term present in Agrotags but the term 'peanuts' is not present; so if a document contains a keyword 'peanuts' it will be mapped to 'groundnuts' term of Agrotags.

2. Similarly, 'Fallow Systems' (termcode-2784) is a term in Agrovoc as well as Agrotags. 'Bush fallowing' (termcode-1160) is a narrower term (NT) of 'Fallow Systems' in Agrovoc but not in Agrotags.

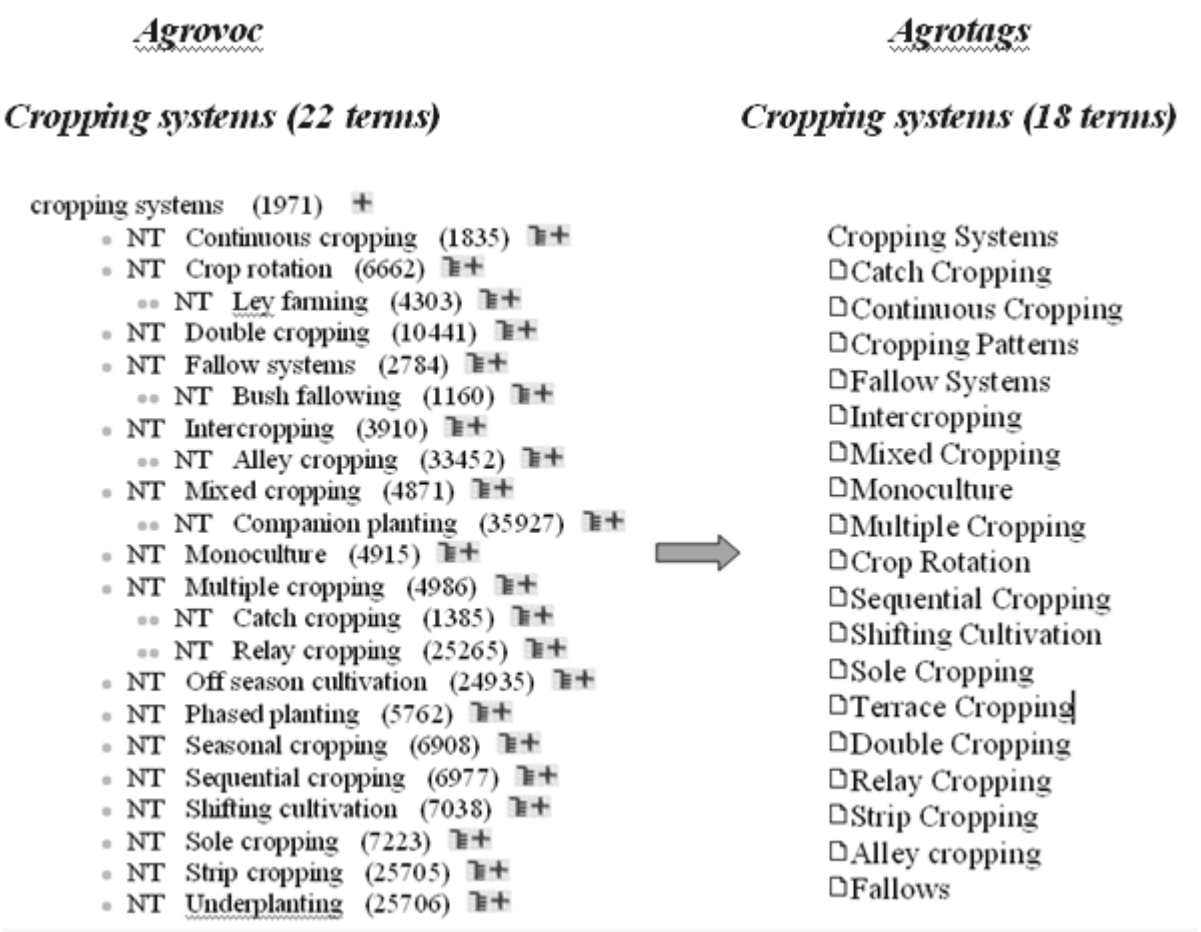

Fig. 1 — Terms at first hierarchical level in Agrovoc were considered in the creation of agrotags hierarchy.

Now if our document contains "bush fallowing" as its candidate term it will be mapped to its broader term, viz. 'Fallow Systems'.

Scientific names and geographical names were also excluded and it was decided to address only crop-related agriculture domain in this version of Agrotags, resulting in the removal of fishery-related terms as well. The concepts selected in Agrotags are hand-picked based on their utility in a tagging scheme as well as their popularity.

Top level terms of Agrotags

The Agrovoc has 25 top level terms whereas Agrotags has 15 top level terms. Agrotags top level terms are not exactly a subset of Agrovoc top level terms but can be termed as broader subset of the overall Agrovoc. The top level terms of Agrotags are listed below:

- Plant Protection
- Plant Production
- Postharvest Management
- Beneficial Insects
- Crop Improvement
- Crops
- Atmospheric Science
- Animal Husbandry
- Environmental Science
- Forestry
- Social Sciences
- Statistics and Experimentation
- Soil Science and Microbiology
- Agricultural Engineering
- Biochemistry and Plant Physiology

Agrovoc to Agrotags term mapping

The broader mapping between Agrovoc and Agrotags terms are indicated in the diagram (Figure 1). The broader and narrower terms (NT) are represented with 'has subclass' relationship. The descriptors and no-descriptors are represented with 'used for' relationship.

Use of Agrotags in OpenAgri

Open Access (OA) repositories have rapidly turned into a global platform for dissemination of the scientific literature[6]. OpenAgri is an open source repository for agricultural research documents. It was developed by IIT Kanpur and is accessible through the web[7]. It has document types like journal articles, conference papers, book chapters, proceedings, preprints, multimedia content, etc. which allows useful metadata to enable searching and retrieval. This repository provides for rich semantic interlinking between document using Agrotags. Documents are also automatically tagged using the Agrotagger algorithm. Agrotagger is an automatic tagging module for agriculture-related contents[8,9]. It identifies the occurrence of Agrovoc terms in the document, replaces them with an equivalent Agrotags term and then chooses the candidate keyword from among them.

Conclusions

In this paper we have described a system for generating keywords for agricultural research documents. We have proposed a new tag set called Agrotags, which is a broad subset of Agrovoc. Agrotags are specially designed with tagging in mind. We are also developing Agrotagger, a software for assigning key phrases automatically from Agrotags. Agrotagger works by recognizing the Agrovoc terms from the document, mapping them to Agrotags terms and using statistical techniques for assigning probabilities as candidate keywords. The whole system has been implemented and deployed as a web service.

Acknowledgement

Financial and technical support and advice from the Indian Council of Agricultural Research (ICAR), New Delhi, India and the Food and Agriculture Organization (FAO), Rome, are gratefully acknowledged.

References

1. Garfield E, Historiographic mapping of knowledge domains literature, *Journal of Information Science*, 30 (2) (2004) 119–145; DOI: 10.1177/0165551504042802.
2. Balaji V, The fate of agriculture. *http://www.india-seminar.com/2009/597/597_v_balaji.htm*
3. The Agropedia, available at *http://agropedia.net*
4. Kaur P, Patwar S, Sylvester A G and Balaji V, Use of semantic Wiki tool to build a repository of reusable information objects in agricultural education and extension: results from a preliminary study, *Paper presented at the Web2fordev Conference, Rome(25-27 September, 2007) , available at http://vasat.icrisat.ac.in/images/New%20Folder/Web%202%20for%20Dev.pdf*
5. Agricultural Information standards(AIMS) *http://aims.fao.org/*
6. Kayvan K, Mahshid A, The citation impact of open access agricultural research: a comparison between OA and non-OA publications, *IFLA World Library and Information Congress: 75th IFLA General Conference and assembly (23-27 August, 2009), http://www.ifla.org/files/hq/papers/ifla75/101-kousha-en.pdf.*
7. OpenAgri, An open access agricultural research repository, *http://agropedia.iitk.ac.in/openaccess/* .
8. Balaji, V, Meeta Bagga Bhatia, Rishi Kumar, Lavanya Kiran Neelam, Sabitha Panja, T.V. Prabhakar, Rahul Samaddar, Bharati Soogareddy, Asil Gerard Sylvester, Vimlesh Yadav, Agrotags – a Tagging Scheme for Agricultural Digital Objects, *Accepted for Publication, 4th Metadata and semantics research conference, Alcala, Spain 20-22 October, 2010.*
9. Rishi Kumar, Automatic keyword extraction using enhanced knowledge models, *Master's Thesis, Department of Computer Science and Engineering, IIT Kanpur, 2010.*