Agrotags – A Tagging Scheme for Agricultural Digital Objects

Venkataraman Balaji[1], Meeta Bagga Bhatia[1], Rishi Kumar[1], Lavanya Kiran Neelam[1], Sabitha Panja[2], Tadinada Vankata Prabhakar[1], Rahul Samaddar[1], Bharati Soogareddy[2], Asil Gerard Sylvester[2], and Vimlesh Yadav[1]

[1] Indian Institute of Technology Kanpur, India
[2] International Crops Research Institute for the Semi-Arid Tropics , Hyderabad, India
{tvp,meeta}@iitk.ac.in

Abstract. Keyword assignment is an important step towards semantic enablement of the web. In this paper we describe a taxonomy called Agrotags which is designed for tagging agriculture documents. Agrotags is a subset of Agrovoc and is much smaller: about 2100 as against 40,000. 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 Agris databases. Further extending the usage of Agrotags emerges the concept of Agrotagger which is a system for automatically generating keywords for agricultural documents. Agrotagger has been built by moving the learning (what keyword to assign) from the example (document) level to the model level. Agrotagger being a pluggable module can act as an add-on to any repository.

Keywords: Agrovoc, Agrotags, Agrotagger, Keyphrase Assignment, Metadata.

1 Introduction

The near absence of agriculture and farming as distinct practices in the world of Web 2.0 has already been pointed out in many instances [1]. International and national level efforts, like the agropedia[2], have initiated strategies and created pathways to address this problem of bringing quality extension materials to the web. These are stored in a reusable fashion thus facilitating reuse in various contexts across diverse delivery mediums.

However, we find that there is no paucity of research reports, papers and documentation related to agricultural research on the web. Many reputed publishing houses hosts many of these articles in their repository. A few of these repositories use various tagging methods to label documents to facilitate ease of 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 infer knowledge from the tags. This greatly limits the participation and availability of the document across a semantic network.

The need for a knowledge model grounded tagging methodology was strongly felt [22]. The combination of advanced tagging, metadata and cross-linking facilitated by controlled ontologies would give raise to a wealth of semantically-linked and relevant

S. Sánchez-Alonso and I.N. Athanasiadis (Eds.): MTSR 2010, CCIS 108, pp. 36–45, 2010.

documents. Many international agricultural thesauri exist like Agrovoc [3], CABI [4], NAL [5] etc. Agrovoc, with its existence since 1976 as a thesaurus and its morph into a full-fledged agricultural ontology in the last decade, was seen as a natural choice as a base set 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. Agrovoc has provided the glue for the semantic inference in this endeavor [9].

ICRISAT(The International Crops Research Institute for the Semi-Arid Tropics) has long been involved with the Agrovoc enrichment together with the FAO(Food and Agriculture Organization of the United Nations) and with and IITK(Indian Institute of Technology Kanpur) maintaining the Hindi version of Agrovoc. ICRISAT has led the revision and refinement of the Agrovoc thesaurus which forms the basis of the Agrovoc 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 more efficient and simpler searching and provide most efficient resources to the user.

Agrotags's pedigree has been Agrovoc - the agricultural thesaurus from FAO. The ongoing efforts to enrich Agrovoc to ontology is widely known (AIMS website) [6]. Agrovoc is also working on mapping onto leading thesauri such as NAL, CABI, etc this provides documents tagged with Agrotags rich interconnection with documents tagged with 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 greater reach.

1.1 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.

Agrovoc contains (as of May 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 IITK decided that a collection of hand-picked terms would go into the creation of a collection of terms for tagging agriculture related documents.

Initially, the top term creation was based on popular thesauri like NAL and CABI, but later it was decided to create a hierarchy rooted in the concepts from the subject categories in Agris database[7], since these seemed to be better suited for indexing . After the top terms were finalized, the team set about creating the hierarchy taking care to retain the intended purpose of Agrotags. Terms were also sourced outside 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 term of Agrovoc top term *methods* with different depth level. However, *outbreeding* and *mass selection* associated to *crop improvement*, *cultivar selection* to *plant production* and *control methods* to *plant protection* top term of Agrotags.

It was felt that we need to include some terms into Agrotags which are not in the existing version of Agrovoc. Agrovoc is a dynamic and evolving ontology which invites new additions and corrections. So we proposed that these new terms be added to Agrovoc (proposal pending), thus conserving the property that Agrotags is a proper sub-set of Agrovoc. These terms were arrived at by examining the manually assigned tags to more than 2000 English language documents in Agris database during the period 2002-2009. In the first version of Agrotags, 15 top level terms were created. The subsequent revisions may refine these classifications.*Plant production, plant protection, crop* improvement etc., formed some of the top-level terms of this kind.

Currently Agrotags are available in English, Hindi and French languages. Telugu and Kannada versions are in progress. Agrotags can be seen at http://agropedia.iitk.ac.in/agrotags_version2/agro_tree.html.

1.2 Criteria of Selection

Only descriptors and more popular terms were selected to create Agrotags from Agrovoc. The non descriptors, 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 like:

'Rice' is a term in Agrovoc (termcode-6599) and has non-descriptor 'paddy' [8]. 'Rice' is a term present in Agrotags but the term 'paddy' is not present so if our document consists of a keyword 'paddy' it will be mapped to 'Rice' term of Agrotags Similarly, 'Organic Wastes' (termcode-35237) is a term in Agrovoc as well as Agrotags. 'Garden Wastes' (termcode-35242) is a narrower term (NT) of 'Organic Wastes' in Agrovoc but not in Agrotags. Now if our document consists of Garden Wastes as its candidate term it will be mapped to its broader term that is 'Organic Wastes'.

Scientific names, geopolitical names were also excluded and it was decided to address only agriculture domain in this edition of Agrotags resulting in the removal of fisheries related terms as well.

To summarize, the following equation describes the relationship between Agrotags and Agrovoc:

Agrotags = Agrovoc - (Non_Descriptor terms**+** Scientific Terms **+** Geopolitical Terms **+** Fisheries)

1.3 Top Level Terms of Agrotags

Agrovoc has 25 top level terms where as Agrotags has 15 Top level terms. Agrotags top level terms are not a subset of Agrovoc top level terms but a subset of the overall Agrovoc(Fig.1).

1.4 Agrovoc to Agrotags Term Mapping

The diagram below (Figure 2) shows the hierarchical structure of the 'Methods' fragment of the Agrovoc ontology. The terms in red are the one included in Agrotags. Relationship information NT: Narrower Term, usedFor: Non-Descriptor. Fig.3 shows a table for mapping between Agrovoc to Agrotags terms.

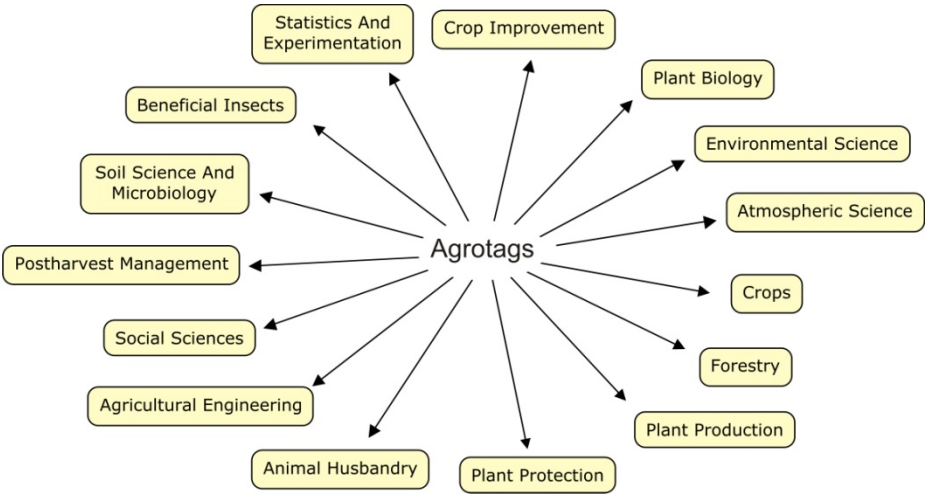

Fig. 1. Agrotags top-level terms

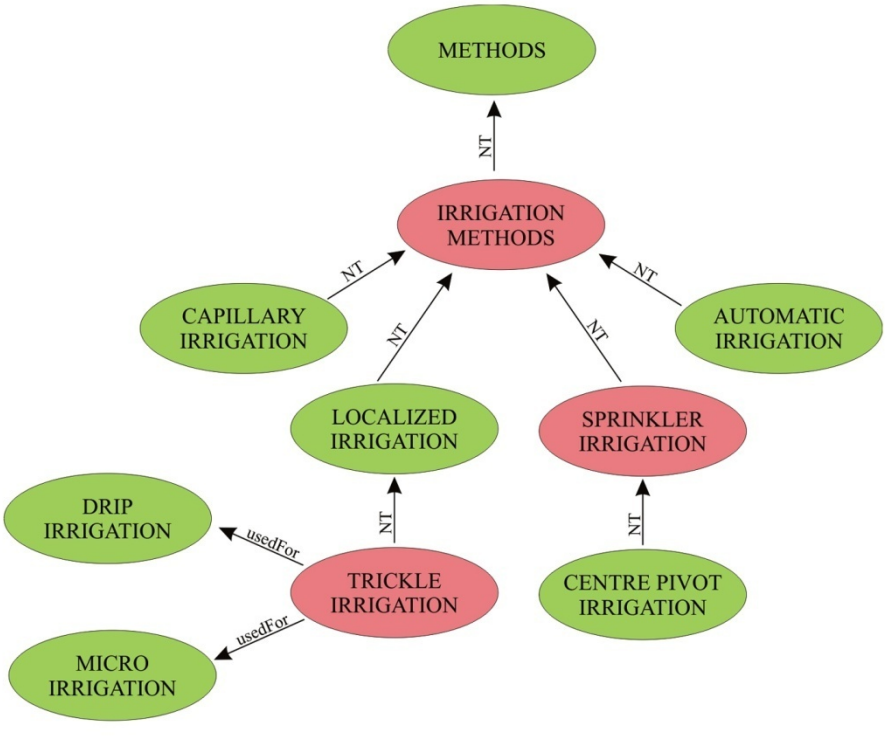

Fig. 2. Agrovoc to Agrotags term mapping

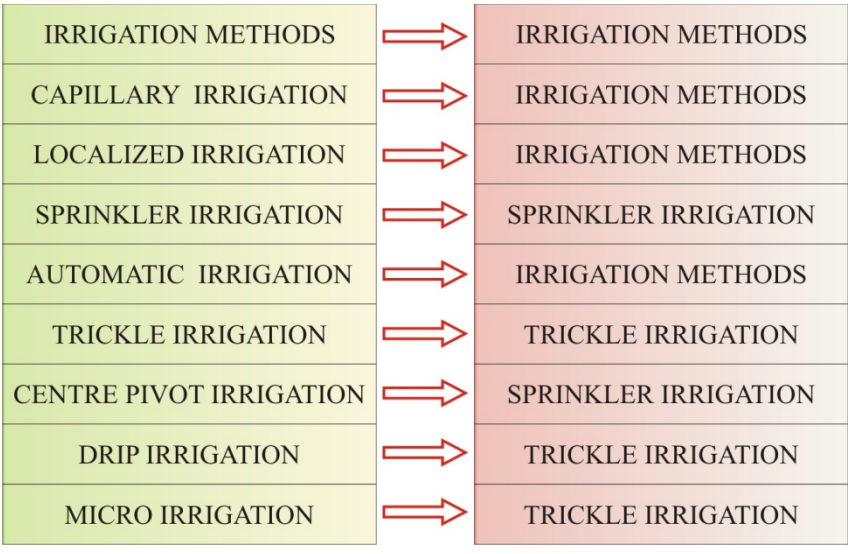

Fig. 3. Term mapping between Agrovoc to Agrotags

Fig. 4. Agrotags, Agrotagger and openagri in joint action

1.5 Use of Agrotags

Agrotags currently are stroed in an internal database format which is used by OpenAgri[10], an open source repository for agricultural documents developed by IIT-Kanpur and ICRISAT. This repository provides for rich semantic interlinking

between document using Agrotags Documents are also automatically tagged using the Agrotagger algorithm (Fig.4). Also refer [18], [19], [20], [21] in Open access context.

2 The Agrotagger

Machines as compared to human give more efficacious result in almost all the domain, but when it comes to natural language understanding, machine driven results can't compete human analysis. But this also has a positive side, extracting a handful of keywords from content potentially seems to be a feasible solution and with that point a pluggable module called Agrotagger is being developed with collaboration of FAO. This module could be used as an add-on to leading repositories such as DSpace and advanced management systems like Drupal and Joomla to automatically tag documents within a controlled vocabulary such as Agrotags. User generated tags together with those that are generated by Agrotagger would help link documents related to agriculture more effectively for faster retrieval and for an enhanced presence in the present flair of the web.

2.1 Need for Agrotagger

With the huge amount of digital documents existing in the internet and their growing panoply with each passing day, keyphrases prove to be an important metadata. Although key phrases can be assigned by the document's author at the time of its creation, the manual process of tagging the documents with keyphrases is not only labor-intensive and time-consuming but also yields poor indexing consistency over the entire document collection.

Indexing a document is not a very new concept indeed- if we take a brief look in the Ancient History, we will find that long back in fourteenth century, the first systematic approach to indexing emerged which was true alphabetical indexing. Later as the technology developed fresh ideas kept coming and alphabetical index became catalogue, catalogue became taxonomy, taxonomy gets converted to thesaurus and then using this vocabulary we get automatically generated keywords from Agrotagger.

Any given document's metadata consists of fields like: author, title, keywords etc. but the most reliable of all is keywords. For example: The title "Options for adaption, though limited do exist" is an article about Marine fisheries from the magazine "The Hindu- Survey of Indian Agriculture 2009". Now the given title has no clue about the actual topic of the article. This is where keywords are crucial.

Automatic keyword assignment has several approaches, primarily keyword assignment from a vocabulary where the candidate keyword is from a standard vocabulary and keyword generation from text where we do not restrict the candidate keyword to a specific vocabulary. These could be rule-based assignments or based on machine learning. Some of the sample rule-based systems are E. Han and G. Karypis [13], L.S. Larkey and W.B. Croft [14], Fabrizio Sebastiani [15]. Eibe Frnak [16], P.D. Turney [17], are based on machine learning.

2.2 Role of Agrotags in Agrotagger

Agrotagger uses Agrotags as candidate key phrases for documents. As explained earlier Agrotags are a proper subset of Agrovoc – Agrovoc has about 40,000 agricultural

concepts and Agrotags has around 2100. The concepts selected in Agrotags are hand-picked based on their utility in a tagging scheme as well as their popularity. Agrotagger 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.

2.3 Workflow in Agrotagger

At the top level, Agrotagger works in three main stages:

Stage 1: Identify all Agrovoc terms in the document – the document now is a bag of Agrovoc terms

Stage 2: For each of these Agrovoc terms, identify an Agrotags term; this reduces the document to a bag of Agrotags terms.

Stage 3: Use statistical techniques to calculate the suitability of these terms for key-phrases

Agrotagger is inspired by an automatic keyphrase extraction algorithm called KEA[11]. Basically the KEA system works by training a classifier (which is done through training the system using large datasets) and keyword assignment using the trained model. Learning through a large corpus is difficult – they are not simply available. We have modified the KEA algorithm by shifting the training from the corpus to the knowledge model level – the Agrovoc to Agrotags mapping is the learning model and has been manually constructed.

After obtaining content bearing terms (by eliminating fluff words and through stemming) we intersect them with Agrovoc terms. The resulting terms are then mapped with their respective Agrotags terms from a pre computed Hash Table. This set of filtered candidate terms are then given as an input to the KEA algorithm.

To extract keyphrases KEA makes use of the following attributes:

- Length of a phrase in words
- Frequency of the words
- Node Degree of the candidate terms
- Occurrence based on location of the terms
- Appearance: Binary Variable to check the presence of the terms.

For more details refer: KEA: Keyword Extraction Algorithm and Rishi Kumar'sThesis [12]

Figure 5, gives the top level workflow in Agrotagger. Stopwords is the name given to words which are filtered out prior to, or after, processing of any selected document. In our case we have identified 262 distinct stop words which are generally articles, pronouns, adverbs, prepositions, conjunctions, consonants, vowels and some unit entities.

2.4 Usage of Agrotagger

It is currently being used by an open access agricultural research repository called openagri. This repository is a open platform to submit any kind of agricultural published material under a single hood, all a user needs is a username and password which is easily attainable by registering into the site. Once a user registers and submits his document, the Agrotagger running in the background automatically generates keywords. See Fig.6 for a sample screen.

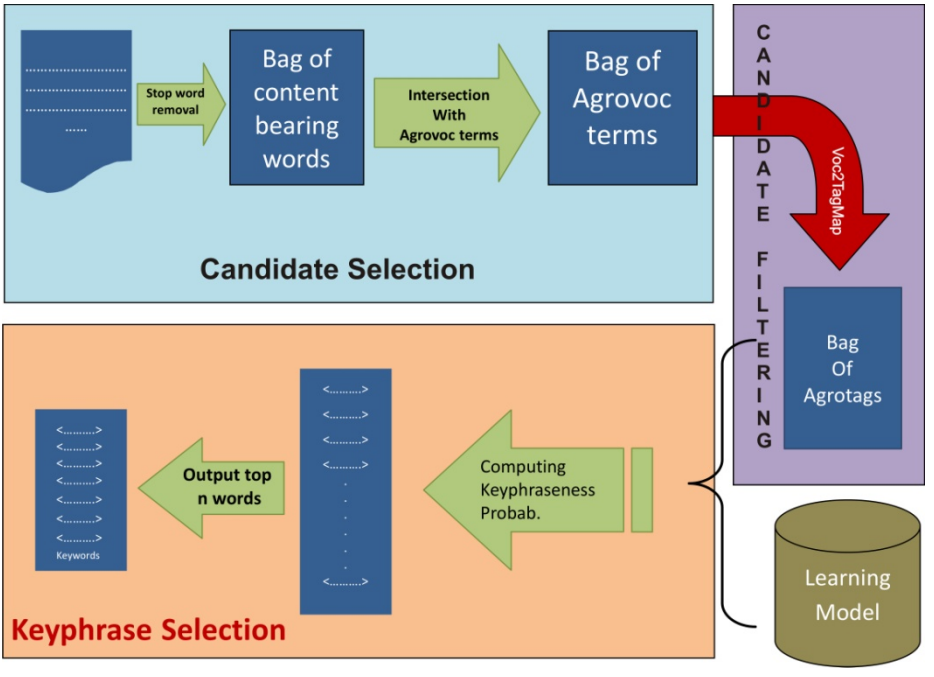

Fig. 5. Workflow of Agrotagger

http://agropedia.iitk.ac.in/auto_tagger/callable_auto_tagger.php

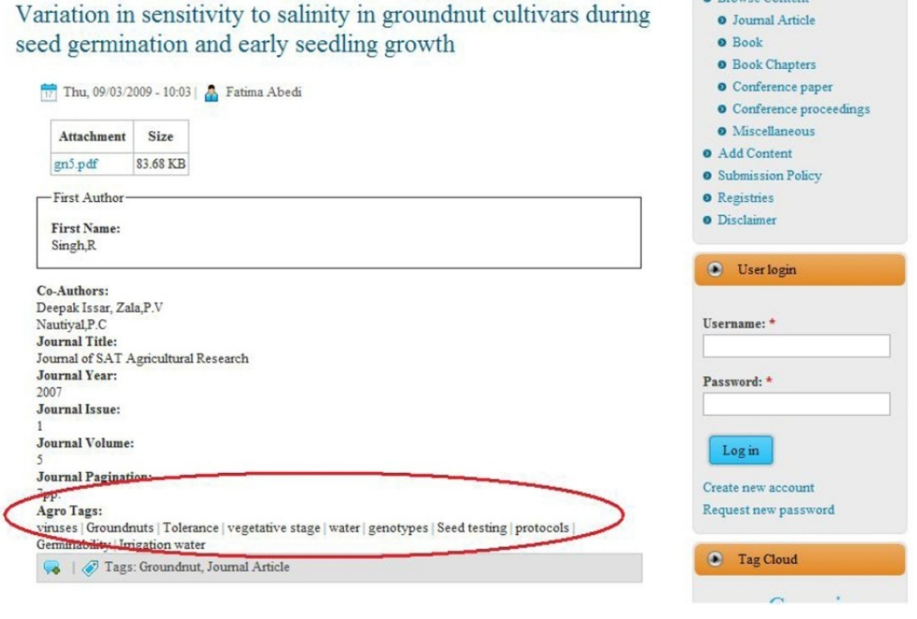

Fig. 6. Document from openagri research repository

Agrotagger is also available as a web service. To automatically get keywords for your agricultural document (as of now only pdf's) go to:

3 Conclusions

In this paper we have described a system for automatically generating keywords for agricultural documents. We propose a new tagset called Agrotags, which is proper-subset of enhanced Agrovoc. Agrotags are specially designed with tagging in mind. Agrotagger is a software for assigning keyphrases automatically from Agrotags. Agrotagger works by recognizing Agrovoc terms from the document, mapping them to Agrotags terms and using statistically techniques for assigning probabilities as candidate keywords. The whole system has been implemented and deployed as a web-service.

Acknowledgement

We gratefully acknowledge the Indian Council of Agricultural Research, New Delhi, India and Food and Agriculture Organization , Rome for their unending Financial, technical and overall support and cooperation.

References

1. Balaji, V.: The fate of agriculture, http://www.india-seminar.com/2009/597/597_v_balaji.htm
2. Agropedia: An agricultural encyclopedia, http://agropedia.net/
3. Agrovoc: A multilingual agricultural thesaurus, http://aims.fao.org/website/Agrovoc-Thesaurus/sub
4. CABI, http://www.cabi.org/
5. National Agricultural Library, http://www.nal.usda.gov
6. Agricultural Information Management Standards, http://aims.fao.org/
7. AGRIS: International Information System for the Agricultural Sciences and Technology, http://agris.fao.org/
8. Agrovoc: A multilingual agricultural thesaurus-Terminology, http://www.fao.org/docrep/008/af234e/af234e02.htm
9. Use of Semantic Wiki Tools to Build a Repository of Reusable Information Objects in Agricultural Education and Extension: results from a preliminary study: Web2ForDev International Conference, Rome (September 25-27, 2007), http://www.web2fordev.net/
10. Openagri: An Open Access Agricultural Research Repository, http://agropedia.iitk.ac.in/openaccess/
11. KEA: Keyword Extraction Algorithm, http://www.nzdl.org/Kea/index_old.html
12. Kumar, R.: Master's Thesis:Automatic Keyword Extraction suing Enhanced Knowledge Models

13. Han, E., Karypis, G.: Centroid-based document classification analysis and experimental result. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 424–431. Springer, Heidelberg (2000)
14. Larkey, L.S., Croft, W.B.: Combining classifiers in text categorization. In: SIGIR, pp. 289–297 (1996)
15. Sebastiani, F.: Machine Learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
16. Frnak, E., Gautwin, C., Manning, C.G.N., Witten, I.H., Paynter, G.W.: Kea: Practical automatic keyphrase extraction. In: Designand Usability of Digital Libraries: Case Studies in the Asia Pacific, pp. 129–152 (2005)
17. Turney, P.D.: Learning to extract keyphrases from text. Technical report, National Research Council, Institute for Information Technology (1999)
18. Kousha, K., Abdoli, M.: The citation impact of Open Access Agricultural Research: a comparison between OA and Non-OA Publications. In: IFLA World Library and Information Congress: 75th IFLA General Conference and Assembly (2009), http://www.ifla.org/files/hq/papers/ifla75/101-kousha-en.pdf
19. Open Access Publishing: Views of Researchers in PublicAgricultural Research Institutions in Zambia by Justin Chisenga and Davy Simumba. Agricultural Information World wide (2009)
20. Gawrylewski, A.: http://www.soros.org/openaccess/read.shtml (2008)
21. John, H., Sheehan, P.: 2006 The Economic Impact of Enhanced Access to Research Findings. CSES Working Paper No. 23. Agricultural Information Worldwide - 2 (2009), http://www.cfses.com/documents/wp23.pdf
22. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering thesauri for new applications: the AGROVOC example. Journal of Digital Information 4(4) (2004), http://journals.tdl.org/jodi/article/viewarticle/112/111