# Automatically Categorizing Metadata Databases into a Categorization Scheme on a Large Scale Web Site

Barbara Richards, Boris Lauser, Anita Liang, Johannes Keizer, Stephen Katz
Food and Agriculture Organization of the UN, Italy
{barbara.richards, boris.lauser, anita.liang, johannes.keizer, stephen.katz}@fao.org

## Abstract

*In order to provide thematic access to the large scale web site of the Food and Agriculture Organization (FAO), we have adapted a scheme used in bibliographical databases to allow subject browsing on FAO's web site. This requires categorizing existing database records across the organizations web sites into the new scheme. An algorithm to automatically categorize these metadata records based on their subject descriptors has been devised.*

## 1. Introduction and background

The Food and Agricultural Organization (FAO) is one of the largest specialized agencies in the United Nations system and the lead agency for agriculture, forestry, fisheries and rural development. Its web site has approximately 40 GB of data.

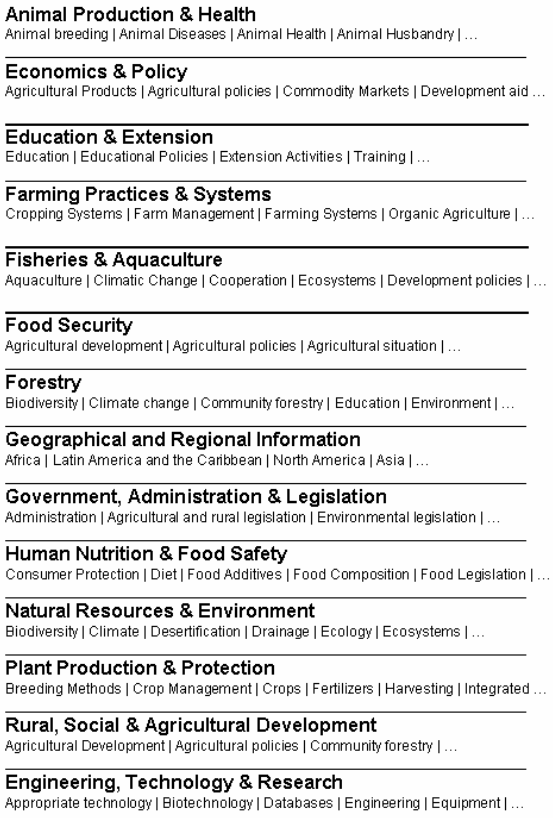
The main tools to provide thematic and subject access to FAO's information in databases across the organization are AGROVOC[1], a multilingual thesaurus, and the AGRIS categorization scheme[2]. Currently, these are used in a number of databases throughout the organization, such as the FAO Library and Documentation database[3], the Document Repository[4] and the FAO Information Finder[5].

Two main problems prevent consistent subject access across the FAO web site: 1) It is difficult to ensure consistent categorization throughout the organization; and 2) the AGRIS categorization scheme is not entirely reflective of the information available throughout FAO, i.e. categories are unequally populated with information.

In order to address the latter problem, a new categorization scheme was developed. This requires converting database records to the new categorization scheme. To address inconsistent categorization and the necessary re-categorization, an algorithm was developed to generate categories from subject descriptors.

## 2. The new Categorization Scheme

Figure 1 shows the 15 main categories and sub-categories.



**Animal Production & Health**
Animal breeding | Animal Diseases | Animal Health | Animal Husbandry | ...

**Economics & Policy**
Agricultural Products | Agricultural policies | Commodity Markets | Development aid ...

**Education & Extension**
Education | Educational Policies | Extension Activities | Training | ...

**Farming Practices & Systems**
Cropping Systems | Farm Management | Farming Systems | Organic Agriculture | ...

**Fisheries & Aquaculture**
Aquaculture | Climatic Change | Cooperation | Ecosystems | Development policies | ...

**Food Security**
Agricultural development | Agricultural policies | Agricultural situation | ...

**Forestry**
Biodiversity | Climate change | Community forestry | Education | Environment | ...

**Geographical and Regional Information**
Africa | Latin America and the Caribbean | North America | Asia | ...

**Government, Administration & Legislation**
Administration | Agricultural and rural legislation | Environmental legislation | ...

**Human Nutrition & Food Safety**
Consumer Protection | Diet | Food Additives | Food Composition | Food Legislation | ...

**Natural Resources & Environment**
Biodiversity | Climate | Desertification | Drainage | Ecology | Ecosystems | ...

**Plant Production & Protection**
Breeding Methods | Crop Management | Crops | Fertilizers | Harvesting | Integrated ...

**Rural, Social & Agricultural Development**
Agricultural Development | Agricultural policies | Community forestry | ...

**Engineering, Technology & Research**
Appropriate technology | Biotechnology | Databases | Engineering | Equipment | ...

**Figure 1**: The FAO category map

### 2.1. Mapping records to categories

We associated the subject descriptors (i.e. keywords taken from the AGROVOC thesaurus) of metadata records to one or two of the 15 main categories. Table 1 shows a small sample of the mapping.

**Table 1**: Descriptor – category mapping

| AGROVOC descriptor | Mapped Main Category |
|---|---|
| Agroforestry | Forestry |
| Agroforestry | Farming Practices & Systems |
| Agronomy | Plant Production & Protection |
| Community involvement | Rural & Social Development |

### 2.2. The Categorization algorithm

[1] http://www.fao.org/agrovoc/
[2] http://www.fao.org/docrep/003/U1808E/U1808E00.htm
[3] http://www4.fao.org/faobib/
[4] http://www.fao.org/documents/
[5] http://infofinder.cgiar.org/

The categorization algorithm is based on the subject descriptors in a metadata record. It assigns the categories based on the mapping described, taking into consideration the number of subject descriptors in the metadata record:

- When there are 4 or less subject index terms, 1 term matching to category generates the category
- When there are 5-10 subject index terms, 2 terms matching to category generates the category
- When there are 11-15 subject index terms, 3 terms matching to category generates the category
- When there are 15 or more subject index terms, 4 terms matching to category generates the category

The following example illustrates how the category for the home page of the Codex Alimentarius Commission[6] would be calculated from its metadata record:

**Table 2**: Mapping of Codex Alimentarius web site subject descriptors to their respective categories

| **AGROVOC terms in metadata** | **Mapped category(s)** |
|---|---|
| Codex alimentarius | Law & Administration + Food Safety |
| Consumer protection | Human Nutrition & Food Safety |
| Food legislation | Law & Administration |
| Food safety | Human Nutrition & Food Safety |
| Foods | Human Nutrition & Food Safety |
| International trade | Economics & Policy |
| Quality | Ignored |
| Standards | Law & Administration |

The above record contains 8 terms. Therefore, only categories associated with at least 2 subject terms will be assigned. The categories assigned by the algorithm are "Law & Administration" and "Food Safety".

### 2.3. Test set

In a first exercise, 242 of FAO's websites were manually categorized by an indexing professional to appear under the 15 main categories presented in Figure 1. The websites are from various departments applying different subject indexing practices and therefore serves as a good test to apply the algorithm. Overall, 386 category assignments were made by the professional indexer, resulting in an average of 1.6 categories per document.

### 2.4. Application of the algorithm and results

Running the algorithm on these records resulted in 408 assigned categories, an average of 1.7 categories per website. 288 of the automatically created category mappings matched exactly with the human categorization, producing a recall of 75%. 120 of the automatically created categorizations did not match with the human approach (a precision of 70%). These unmatched 120 categorizations were then examined by a second professional indexer uninvolved in creating the initial human categorization. The majority have still been found to be acceptable.

### 2.5. Analysis of results

A number of indexing consistency studies show that consistency between indexers varies a great deal, with ranges between 4 and 82 percent. [2]. In a study of subject cataloging of 956 academic journal articles in the field of Library and Information Science, the average consistency of main categories assigned was 82.46%, and that of subcategories was 69.82% [1].

Considering this indexer-indexer inconsistency across human indexers, the 70% consistency with professional human categorization this algorithm achieved shows considerable promise.

However, aside from the indexer inconsistency there are some further limitations of the algorithm to consider:

- The relatively small size of the test set
- Further experimentation with the algorithm parameters might lead to different results
- The algorithm is limited to quality of the subject descriptor indexing by different human indexers
- Having a categorization scheme with significantly higher number of categories might reduce the consistency found with 15 categories

Overall, we have developed a more user-friendly, content driven categorization scheme to access the FAO web site that distributes information more evenly than the previous scheme. Inconsistent decentralized metadata creation has also been addressed by introducing an automatic mapping algorithm with promising results.

Further work includes testing the algorithm with other categorization schemes, larger test sets and the support of consistent, high-quality meta-data creation by both professional and non-professional indexers.

## References

[1] Chen, K.; Lo. S.; Lin, C. (2002). The Investigation of the Consistency of Subject Cataloging for Academic Journal Articles of Library and Information Science.

[2] Yaşar Tonta (1991). A study of indexing consistency between Library of Congress and British Library catalogers, Library Resources & Technical Services 35(2): 177-185. Retrieved 15 May, 2003 from Hacettepe University: http://yunus.hacettepe.edu.tr/~tonta/yayinlar/indxcons.htm

[6] http://www.codexalimentarius.net/