Lessons learned on data discovery, integration and ingestion in AGRIS

Fabrizio Celli (FAO)
DCMI Virtual 2020
22 September 2020
The Food and Agriculture Organization (FAO) is a specialized agency of the United Nations that leads international efforts to defeat hunger and improve nutrition and food security.

It was founded in October 1945.

The FAO is headquartered in Rome, Italy and maintains regional and field offices around the world, operating in over 130 countries.
Initiative set up by FAO in 1974 to make information on agriculture research globally available.

A collection of multilingual bibliographic metadata on agricultural research

A network of nearly 450 data providers from 150 countries

https://agris.fao.org
The AGRIS Network

12 million bibliographic records
3.4 million full-text links
Accessed from 200 Countries and Territories

Records by
434 Data providers
From about 150 Countries
Available in up to 90 Languages
Originally, AGRIS centers were assigned by governments to collect all the scientific production in the country and to send it to AGRIS.

From 2005, AGRIS accepts data also from institutional repositories, journal publishers and aggregators.

With the evolution of technology and the growth of **open access institutional repositories**, AGRIS has improved its methods for harvesting, processing and indexing metadata.
Challenges

Integration of new data in AGRIS
- Variety of metadata formats
- Variety of standards
- Different levels of metadata quality

Automatic ingestion from web APIs
- Understand the relevance of high-volume data (data discovery)
- Content classification and data integration
AGRIS accepts the most common XML metadata formats such as MODS, Crossref, DOAJ, EndNote, MARC21, METS, Simple DC, PubMed and AGRIS AP

The data is curated and converted prior to the AGRIS indexing

The AGRIS team highly recommends to consider **LODE-BD Recommendations 2.0** in order to learn about different metadata terms that can be used to describe properties included in the record
Initial phase: manual validation

Data Collection

National Libraries
Journal Publishers
Institutional Repositories
Aggregators

Metadata validation

Data Processing

Data Publication
In the digital era, many institutions and organizations expose the data on the web.

Big volumes of data from heterogeneous sources raise problems of relevance, data classification, data standardization, data validation, and data provenance.

Data relevance and data classification require new solutions.
Controlled vocabulary covering all areas of interest of FAO, translated into 39 languages

Curated and multilingual list of related contents

It can help with data discovery and classification
The problem of data relevance refers to the ability of harvesting only data that belong to the AGRIS domain.

Data is not always classified, or the classification is very often poor.

The AGRIS solution: machine learning using data already available in AGRIS and the richness of AGROVOC.
AGRIS relies on AGROVOC to enable multilingual search and to connect the data (internally and to external data).

Being able to classify and tag metadata with AGROVOC is important to enrich the semantics of AGRIS content.

The AGRIS solution: machine learning using AGROVOC and natural language processing techniques.
Thank you!

AGRIS@fao.org
http://agris.fao.org