What is your idea?

Many new mobile applications try to reach farmers with tailored, text-based content to improve decision-making. Barriers remain, including low levels of literacy and SMS spam. We have observed, however, that farmers are especially keen on audiovisual content and that ownership of smartphones is generally very high among extension agents and increasing among farmers. The "How to …" category on YouTube is very popular. YouTube has millions of videos with agricultural content. A single search on "How to grow potatoes" returns close to 3 million hits. But finding a relevant video in this vast source of information is problematic. Sophisticated searches require high literacy skills and it is difficult for farmers to assess the quality and impartiality of videos.

We will transform the way farmers and extension agents approach audiovisual information by creating a service for accessing agricultural videos adapted to their needs. The solution will allow searching by voice, make use of relevant contextual information on the users' conditions and will automatically formulate simple, relevant questions to the user to search in a huge existing stock of agricultural information available through YouTube.

What is unconventional in our solution is: (1) it discloses an important unstructured data source, YouTube, rather than focusing on structured data or creating new primary data (2) it is adapted to farmers' and extension agents' capacities by using a combination of voice-based search and cognitive computing, in contrast with text-based or low-interaction solutions.

Our hypothesis is that existing content on YouTube, when enhanced in the way described above, can be used much more intensively than it has been done so far and complement agricultural extension. Success for our project means that farmers consume more videos that help them to take farming decisions. Our project will succeed because: (1) the solution build on an already available, major big data source with proven value to extension agents and farmers, (2) the solution is highly tailored to farmers' capacities and needs in terms of channels, formats and interaction.

How will you pilot it?

The Information Services team of Bioversity (Carlos Quiros, Jacob van Etten et al.) will co-develop the algorithms to process the YouTube videos and form the knowledge bank for IBM Watson (cognitive computing solution) working with IBM Costa Rica (where the Watson Latin-American team is based). We will work with Bioversity's Ontologies group (Elizabeth Arnaud) and Agroecology team (Simon Attwood) to integrate ontologies and data standards into Watson. We will also develop ways for integrating geographic information. Finally, we will use a user-centered design approach and partner with Centros Agricolas Cantonales (sub-province farmers' organizations supported by the Costa Rican Ministry of Agriculture) of Siquirres and Nicoya to co-develop and test different prototypes of the user interface of the platform and best ways to train the IBM Watson cognitive computing algorithms.

CGIAR Platform for Big Data in Agriculture
Inspire Challenge Proposal

We will start our work by developing the algorithms for harvesting the information from YouTube, move this into IBM Watson and connect it with ontologies and geographic information. We will run a series of learning exercises to start training Watson in this new area (determining relevance and quality). Then we will develop ways for connecting Telegram with Watson and start designing the prototypes of the user interface. Through user-centred interaction design we will test and refine prototypes with groups of farmers in Siquirres and Nicoya. The project will develop the following new technologies: (1) massive text harvester from YouTube videos, using Hadoop for efficiency (2) knowledge bank in Watson, (3) algorithms to connect unstructured data with structured data including ontologies and GIS, (4) platform to support user interaction using voice on mobile devices.

The technical work will be done by staff of Bioversity's Costa Rican office under the supervision of Carlos Quiros and Jacob van Etten. The Costa Rican team already works with Hadoop and has the expertise to make the Telegram application and the necessary algorithms to interact with Watson. This will consume 40% of the budget. Another 40% will be spent on working with IBM on creating the knowledge bank and co-developing the algorithms to connect ontologies and GIS data. The fact that a IBM Watson team is located in Costa Rica will make these interactions highly efficient. The remaining 20% will be allocated to our national partners to help us in co-develop and test the prototypes with the farmers. Again, doing this locally in Costa Rica will save resources. The time will be spread 70% in research and development and 30% in testing and feedback.

We will generate several terabytes of data, specifically text extracted from voice in YouTube videos and metadata describing this dataset, linked to ontologies. Also, we will create databases for Watson to assess the relevance of videos, using existing environmental and socio-economic data. The system itself will also quickly generate much data "exhaust" of farmers and extension agents interacting with the Watson-based search system. These unique data will be stored and will serve to analyze farmers' emerging information needs, and can potentially feed early warning systems.

Developing this solution will create interesting perspectives for further innovation. The next steps would be: (1) test the solution under a wider range of biophysical and socio-economic conditions, (2) integrate the solution into the national extension system of Costa Rica (Bioversity has a strong partnership with the extension system, see https://goo.gl/cGuiy1), (3) piloting and promoting in other countries and regions, based on existing Bioversity-led projects (East Africa, India), and in partnership with GFAR and other organizations (4) integrate other sources of unstructured and structured data, (5) develop new services around it, such as periodic reports based on the search data and linkages with input suppliers, (6) develop a business model that could sustain the platform (monetizing advertisement in YouTube, links with input suppliers), (7) encourage targeted, demand-led content creation to fill the information gaps that become evident.