The criteria for collection, validation and management of data about agricultural innovations for smallholders and family farms in the FAO-REU region

Technical paper prepared in connection with the Regional expert consultation on knowledge sharing for agricultural innovations applicable for smallholders and family farmers in Europe and Central Asia

Gödöllő, Hungary, 10-13 September 2018

Laszlo Gabor Papocsi

2018

V.1.0.

Contents

Summary

The word „criteria" means a standard of judgment; a rule or principle for evaluating or testing something (source: dictionary.com). The study about the criteria for data validation and management in the context of the innovation platform therefore requires considerations about objectives aspects which make it possible to clearly decide about the outcome of the examination of a subject.

On one hand, content of a proposed innovative solution should be explicitly evaluable by a set of rules and conditions, that is to be genuinely developed and used by experts of the content domain. On the other hand, many methodologies exist for technical validation of the data, including schema structure, constraints, restrictions, simple and complex rules, etc. So what should be highlighted here is the necessity of appreciating and – as much as possible – using standards and controlled data schemas with different scopes, including technical ones like for example those from the W3C (Word Wide Web Consortium) and especially professional requirements and recommendations like from FAO, ITU, DC and several others.

To be able to set up clear and transparent criterion for data collection on innovation examples for small holders and family farmers, the most required components are:

- guideline about the evaluation of the innovation content, especially its meaningfulness and usefulness,
- data collection template (form) with
    - section using standard metadata schemes, controlled vocabularies and data lists,
    - section with proprietary fields, guidelines provided.

In this paper we provide context of the related information management aspects in the 1st chapter, especially focusing on standards, metadata, open data and interoperability, which are respected as main criterion for professional implementation of the proposed tasks.

In the 2nd chapter data validation and management criteria is discussed in the light of a concrete process workflow that is being designed for the innovation platform. In accordance with the outcomes of the "Regional Expert Consultation on Knowledge Sharing for Agricultural Innovations Applicable For Smallholders" in Hungary in 2018, it is proposed to go beyond the initial task of collecting innovation examples, by accessing information also from the field, gathering problems and needs from smallholders and family farmers, and trying to match – pair and connect - offered innovations with the problems, for solutions.

1. Background on data management

Data management refers to the management of information and data by an organization for secure and structured access and storage. Data management includes a variety of techniques that facilitate and ensure data control and data flow from creation to processing, use and deletion. Data management is implemented through a coherent infrastructure of technology resources and a framework that defines the administrative processes used throughout the lifecycle of the data.

Data validation is a process that ensures the provision of clean and clear data for the programs, applications, and services that use that data. It checks the integrity and validity of data entered into various software and its components. The data validation ensures that the data meets the requirements and quality standards. Data validation is also referred to as input validation. Data validation helps to ensure that the data sent to the connected applications is complete, accurate, secure, and consistent. Validation rules are generally defined in data dictionary or are implemented by data validation software. Types of data validation include code validation, data type validation, data area validation, restriction check and structured validation.

Source: Techopedia

Standards

E-agriculture standards and interoperability components are needed to enable consistent and accurate collection and exchange of agricultural information across geographic and agricultural sector boundaries. Without these components, agricultural information would be susceptible to misinterpretation and difficult to share, due to incompatibilities in data structures and terminologies.

Source: e-Agriculture Strategy Guide

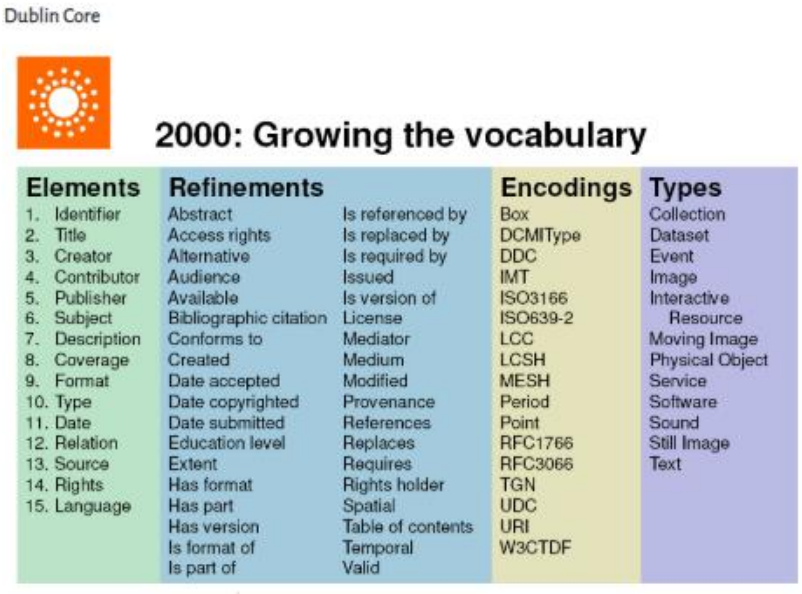
Components	Description	Examples
Data structure standards	These standards govern the way agricultural data sets are stored, using consistent data structures. Data can then be presented with consistency in software applications, to ensure information is neither misinterpreted nor overlooked.	• FAO's Agricultural Information Management Standards (AIMS) supports standards, technology and good practices for open access and open data in the agricultural domain; • Geospatial and sensor data; • Metadata standards, such as Meaningful Bibliographic M etadata (M2B); • Data set compatibility for cross-platform sharing; and • Open data access.
Content quality standards	These standards govern the way that agricultural content is controlled for quality and accuracy.	• Although not a government standard, the GSMA's Agri Guidelines for creating agricultural VAS content are a relevant example; and • Direct2Farm content management guidelines.
Common terminologies	These enable information that is communicated electronically to make use of a common language across e-agriculture platforms for consistency. A localized thesaurus of agricultural terminologies is critical for localization and portability of content across a country/region.	• Agricultural terminology standards, such as AGROVOC.
Secure messaging standards (where necessary)	These are for the secure transmission and delivery of messages and the appropriate authentication of the message receiver, to ensure that information is securely transmitted and delivered to the correct recipient.	• Security standards; • Network and Interoperability standards; • Cloud security standards. For example, ITU-T X Series and Y Series recommendations.
Service interoperability	These define the requirements necessary to conduct various services - such as transactions, information search - across platforms.	• Platform-level interconnectivity; and • Inter-Cloud interoperability. • Financial services interoperability. For example, ITU-T X Series and Y Series recommendations.

Metadata

Metadata can be used to describe the content and properties of a digital object, such as a document, an image, video, audio, website, database, etc. The most widely known and used metadata schemes include: the Dublin Core (DC); the Metadata Object Description Schema (MODS); Virtual Open Access Agriculture and Aquaculture Repository Metadata Application Profile (VOA3R AP); and the AGROVOC thesaurus.

The Dublin Core (DC) is a metadata format that was primarily created for the sake of simple and general web resources descriptions by authors themselves.

The fifteen element "Dublin Core" is part of a larger set of metadata vocabularies and technical specifications maintained by the Dublin Core Metadata Initiative (DCMI). The full set of vocabularies, DCMI Metadata Terms [DCMI-TERMS], also includes resource classes, for example the DCMI Type Vocabulary [DCMI-TYPE], vocabulary encoding and syntax encoding schemes. The terms in DCMI vocabularies can be used in combination with terms from other compatible vocabularies.



Source: HLWIKI International

The original set of 15 metadata elements was extended and refined within the Open Archive Initiative – Protocol for Metadata Harvesting (OAIPMH) (Open Archive Initiative, 2008).

One of the most suitable metadata formats for agriculture is the VOA3R AP. It is partially based on the DC but combined with the AGROVOC thesaurus. As a result, an effective description, availability and automatic data exchange between and among local and central repositories can be attained.

The list of VOA3R Metadata AP elements:

Mandatory	Highly recommended	Recommended	Optional
Title	creator	description	alternativeTitle
Date	contributor	bibliographicCitation	abstract
language	publisher	accessRights	relation
Type	identifier	Licence	conformsTo
Name	format	Rights	references
	isShownBy	reviewStatus	isReferencedBy
	isShownAt	publicationStatus	hasPart
	subject	hasMetametadata	isPartOf
	firstName	personalMailbox	hasVersion
	lastName	objectOfInterest	isVersionOf
		variable	hasTranslation
		Method	isTranslationOf
		protocol	
		instrument	
		techniques	

Source: Simek, 2013

The Agricultural Information Management Standards (AIMS) Platform

AIMS is a platform for accessing and discussing standards for information management in agriculture, of for tools and methods that connect information professionals worldwide to build a global community of practice. AIMS supports a number of projects and initiatives relevant to semantics that: facilitate the provision and exchange of qualitative and interoperable datasets; improve knowledge sharing and reuse; create new collaborative links in the semantic (in agriculture and beyond) ecosystem; and contributes to sustainable agricultural development. The strategies developed, promoted and supported by the AIMS community - to support effective data, information and knowledge management and exchange - focus on:

- descriptive metadata with open-defined and formatted semantics to support use cases beyond bibliographic indexing;
- open vocabularies, concept systems and other knowledge organization systems (KOS) - example see below AGROVOC;
- Semantic Web models and tools, including linked open data (LOD);
- recommendations and best practices for (meta)data publishing and usage, such as AgMES;
- open (and widely used and globally significant) standards and techniques to facilitate mixing and matching data from different distributed infrastructures and resources.

The Agricultural Metadata Element Set (AgMES)

AgMES aims to capture issues of semantic standards in agriculture in terms of description, resource discovery, interoperability and data exchange for different types of information resources. AgMES as a namespace (an abstract container that holds a logical grouping of unique identifiers) should include agriculture specific extensions for terms and refinements from established standard metadata namespaces such as Dublin Core, AGLS, etc. It can be used to attach metadata to document-like information objects, such as publications, articles, books, websites, papers, etc. in the field of agriculture in conjunction with the above-mentioned standard namespaces.

Source: FAO AIMS

AGROVOC Vocabulary

AGROVOC is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment etc. It is published by FAO and edited by a community of experts and editors comprising librarians, terminologists, information managers and software developers.

The vocabulary consists of over 35,000 concepts with up to 40,000 terms in 29 different languages - of different coverage (see SKOSMOS). AGROVOC is made available by FAO as an RDF/SKOS-XL concept scheme - which is a data model for structured controlled vocabularies - and published as a linked data set aligned to 18 other vocabularies.

The AGROVOC thesaurus schema employs three levels of representation:

- concepts represent abstract meanings and are often identified by URIs, e.g. corn as a cereal is identified by „Concept12332“,
- terms are language-specific forms e.g. corn, maïs, or maize
- terms integrating special variants, such as spelling variants, singular or plural forms, e.g. hen, hens, cow or cows.

This is how the abstract concepts/terms and the concrete meanings are related. The AGROVOC is therefore suitable for the description of research papers, information or news in the agrarian sector - Agricultural Information Management Standards.

AGRIS

FAO - AGRIS (International Information System for the Agricultural Science and Technology) is a free of charge service that provides access and visibility to bibliographic data on research papers, reports, multimedia material, grey literature and other content types in agricultural and related sciences. AGRIS is using AGROVOC vocabulary.

Source: FAO AIMS

Open and Linked Data

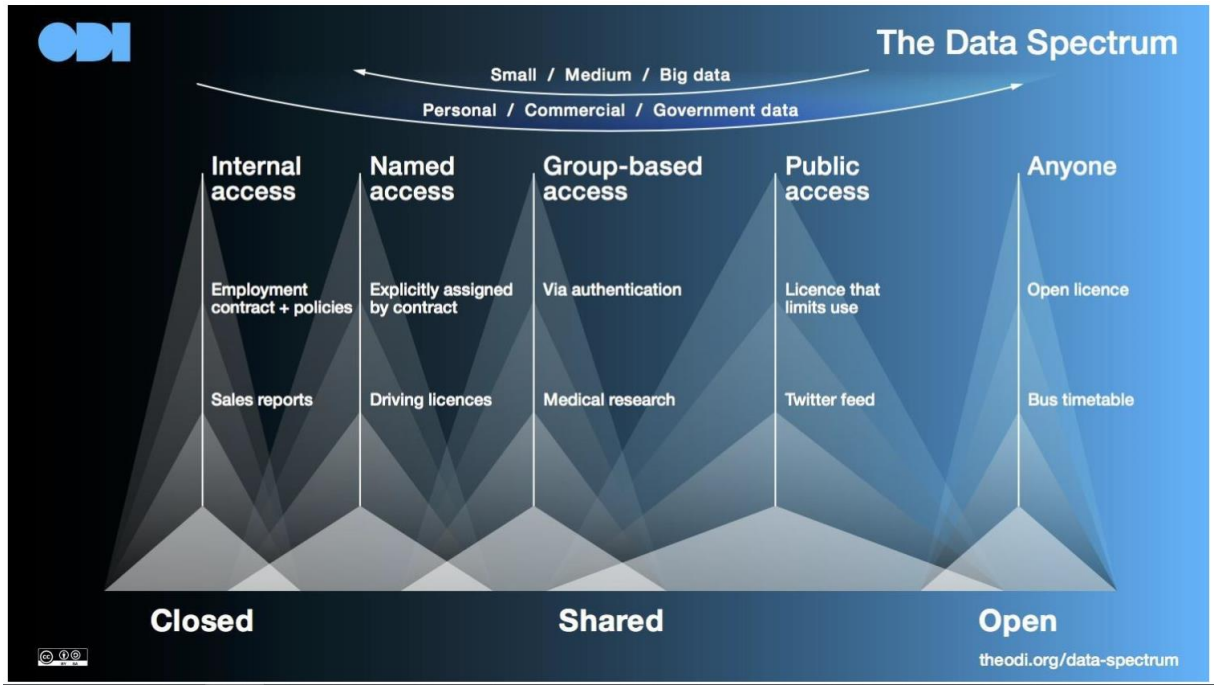
Open data is data that can be freely used, reused (modified) and redistributed (shared) by anyone (Open Knowledge International).

Main criteria of open data:

- Availability and access: The data must be available as a whole with no more than a reasonable reproduction effort, preferably by downloading over the Internet. The data must also be available in a convenient and modifiable form.
- Reuse and redistribution: The data must be provided under conditions that allow for reuse and redistribution, including intermixing with other data sets.
- Universal Participation: Everyone must be able to use, reuse and redistribute - there should be no discrimination between fields of action or between individuals or groups. For example, "non-commercial" restrictions that would prevent "commercial" use or restrictions on use for specific purposes (e.g. education only) are not allowed (Open Knowledge International).

For data to be considered open, it must be:

- accessible, which usually means published on the internet
- available in a machine-readable format
- with a license that allows anyone to access, use and share them - commercial and non-commercial.



Source: The Data Spectrum by the ODI licensed under CC BY

The data spectrum in the figure above, developed by The Open Data Institute (ODI), illustrates the degree of openness of data and helps users to understand the language of the data (the ODI).

Many individuals and organizations collect a wide range of different types of data to perform their tasks. The government is particularly important in this regard, both because of the quantity and centrality of the data it collects, and because most of this government data is public data by law and therefore can be made open and usable for others (Open Knowledge International).

There are many types of open data that have potential uses and applications:

- Culture: Data on cultural works and artifacts - such as titles and authors - generally collected and kept by galleries, libraries, archives and museums
- Science: data produced in scientific research, from astronomy to zoology
- Finance: data such as government accounts (expenditure and revenue) and information about financial markets (equities, stocks, bonds, etc.)
- Statistics: data produced by statistical offices, such as the census and the main socio-economic indicators
- Weather: many types of information to understand and predict weather and climate
- Environment: Information related to the natural environment, such as the presence and level of pollutants, the quality of rivers and seas (Open Knowledge International).

FAIR Data

In 2016, a Nature article "FAIR Guiding Principles for Scientific Data Management and Stewardship" launched the FAIR concept.

FAIR stands for Findable, Accessible, Interoperable, Re-usable principles.

The principles of FAIR Data serve as an international guideline for high quality data management. In the FAIR principles, we use the term "(meta) data" in cases where the principle should apply to both metadata and data.

Although Open Data and FAIR Data are different, they can be overlapping concepts; FAIR data does not automatically mean that it needs to be accessible - for example, sensitive data may have access restrictions.

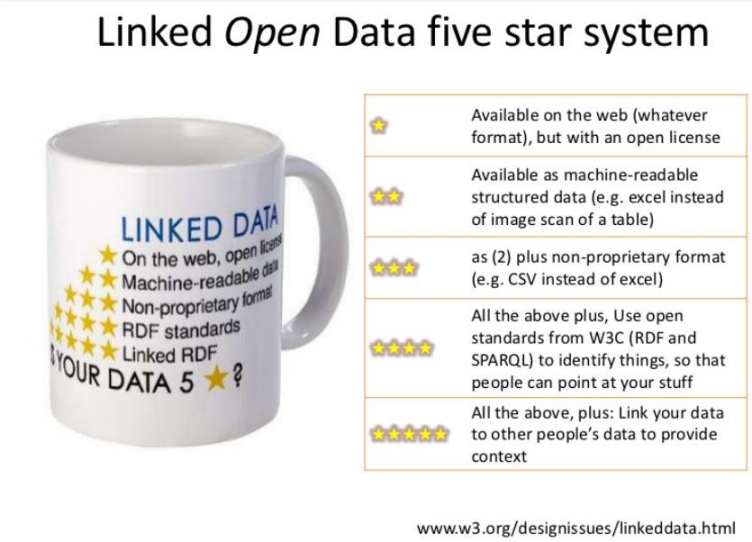
Linked open data

Linked Data is a method of publishing structured data so that it can be interlinked and become more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. (Source: Wikipedia)

Linked Open Data (LOD) is a highly efficient blend of Linked Data and Open Data, being both linked and open source at the same time.

LOD can connect isolated systems with different formats and reduce the obstacles between different sources. It can support the extension of data schemes and updates of the separate data sets without problems of interoperability. It also makes searching complex data easier and more efficient.

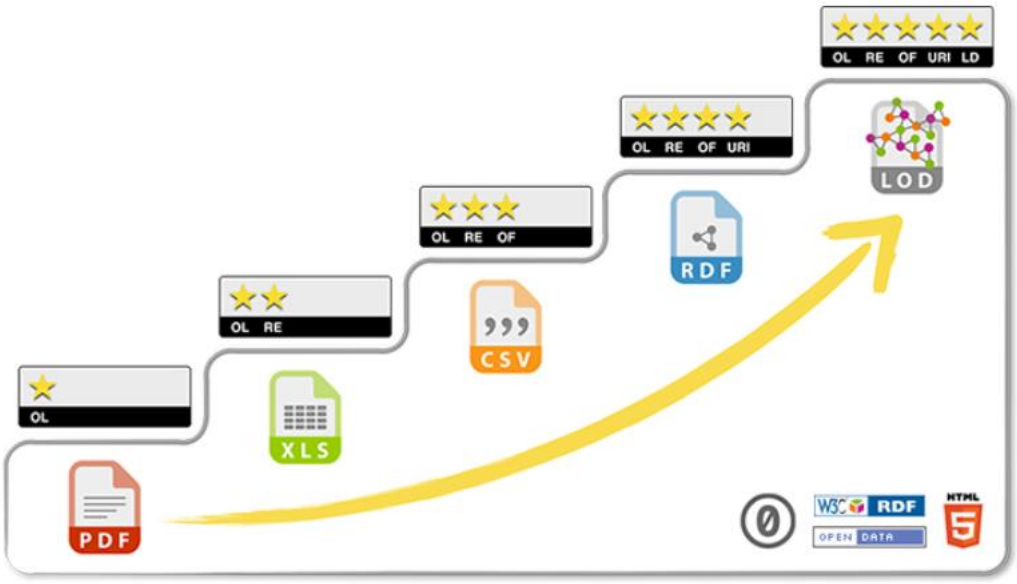
The 5 stars of open data:



To score the maximum five stars, data must (1) be available on the Web under an open license, (2) be in the form of structured data, (3) be in a non-proprietary file format, (4) use URIs as its identifiers (see also RDF), (5) include links to other data sources (see linked data). To score 3 stars, it must satisfy all of (1)-(3), etc.

Source: https://www.w3.org/DesignIssues/LinkedData.html

The 5 starts system explained by file type examples:



Source: http://opendatahandbook.org/

Interoperability

The most commonly used definition of interoperability is: 'the ability of a system or a product to work with other systems or products without special effort on the part of the user'.

For agricultural context, the CIARD community (Coherence in Information for Agricultural Research for Development) defined interoperability for agricultural data as "a feature of datasets … whereby data can be easily retrieved, processed, reused, and re-packaged ('operated') by other systems."

Interoperability can be achieved at different levels, for example so called Foundational (such as transmission protocols), Structural (defining formats and syntax of data exchange) and Semantic Interoperability.

Semantic interoperability which provides interoperability at the highest level, implies the ability of two or more systems or elements to exchange information and to use the information that has been exchanged. Semantic interoperability uses both the structuring of the data exchange and the encoding of the data including vocabulary, so that the systems can interpret the data.

At a structural interoperability level machines understand what different elements are (and their mutual structural relationship), but with semantic interoperability, they also understand the meaning of these elements and can process them with semantic-capable tools to effect advanced deductions.

Semantics for Interoperability of Agricultural Data

Interoperability, the ability of reusing the data produced by others in your own information system, or vice versa, largely depends on how well and explicitly the 'meaning' of the data is described – semantic interoperability. There are three initiatives within the AGRISEMANTICS platform for interoperability between terminologies in agriculture:

a) Agrisemantics working group within the Research Data Alliance develops a set of recommendations for components supporting semantics.
b) GACS Working Group, a project with FAO, CABI, NAL, working to identify a set of concepts common to their three thesauri ("concept schemes", in SKOS term). The output is a concept scheme in beta version.
c) GACS working group is forming a new working group under the umbrella of GODAN. Goal of the group is to enable semantic interoperability of agricultural data, building on the experience of the previous edition of the GACS working group.

CIARD Ring

The CIARD Routemap to Information Nodes and Gateways (RING) is a project implemented within the Coherence in Information for Agricultural Research for Development (CIARD) initiative and is led by the Global Forumon Agricultural Research (GFAR). The RING is a global directory of datasets and data services for the agri-food sector. It is the principal tool created through the CIARD initiative to allow information providers to register their services and datasets in various categories and so facilitate the discovery of sources of agriculture-related information across the world. The RING aims to provide an infrastructure to improve the accessibility of the outputs of agricultural research and of information relevant to ARD management.

Functions of the RING:

- to provide a map of accessible information sources with instructions on how they can be used effectively;
- to provide a dataset sharing platform for the agri-food sector;

- to federate metadata from existing sources whenever possible and alternatively allow for manual submission and curation;
- to provide examples of services that show good practices on implementing "interoperability";
- to clarify the level and mode of interoperability of information sources;

Source:http://ring.ciard.net/sites/default/files/RING-handbook-updated-2017-09-10.pdf

Farm machines

One of the biggest problems farmers face is the interoperability of farming equipment due to different digital standards. This lack of interoperability is not only obstructing the adoption of new IT (Internet of Things) technologies and slowing down their growth in Europe, it also inhibits the gain of production efficiency through smart farming methods. The IOF2020 project aims to integrate different machine communication standards to unlock the potential of efficient machine-to-machine communication and data sharing between machines and management information systems.

About APIs, Web Services

An application programming interface (API) is a set of protocols for building software. A Web API is an application programming interface for either a web server or a web browser.

Good examples

- IrriSAT - http://www.agriteach.hu/en/content/irrisat
- Agro Api - http://www.agriteach.hu/en/content/agro-api
- see other examples on AgriTeach portal

File level interoperability - file formats and conversions

Many information systems enable the download of user defined data in different file formats, and also several make it possible for the users to upload files from their own computer to the same or other online systems. Often file formats used are not the same, so there is a need to convert the data from one file format to the other, in order to make data usable from one system to the other. Most typical file formats used in agricultural digital applications:

- TXT - this is the simple plain text file format, usually to be opened by Notepad under Windows
- CSV - TXT file with values - representing columns in a table - separated by commas (or semicolons), often used to download or upload tabular data
- XLS - similar tabular (spreadsheet) data like CSV, but Microsoft's own Office format, with added complexity
- PDF -typically used to represent printed, finalized, submitted version of some process, or also for filling out offline e-forms

- HTML - usually these are the content pages of the world wide web, can be directly edited and generated by plain html editors or rich text editors and rendered by Content Management Systems (CMS)
- XML - Extensible Markup Language (XML) is a text format that mainly serves the exchange of a wide variety of data on the Web and elsewhere, for instance in web services, M2M communication, e-Government submissions, etc. It is a tag-based structure (similar to html) with extensible nested elements and attributes.
- XSD - is an XML Schema that describes the structure of an XML document, often used in e-Government services to publish the basic structure and rules of documents and forms
- SHP - The shapefile format is a popular geospatial vector data format for geographic information system (GIS) software
- KML - XML format for Google Maps or Earth, used to display geographic data.
- JSON (JavaScript Object Notation) is a lightweight data-interchange format that is easy for humans to read and write, and for machines to parse and generate, based on a subset of JavaScript Programming Language
- RDF - It is a standard model for data interchange on the Web having features that facilitate data mapping even if the merged schemas are different, it also supports changing schemas over time without requiring connected schemas and user applications to be changed.

2. Information management within the proposed Small Holder Innovation Partnership (SHIP) Platform

In this chapter we provide practical recommendations on the implementation of the proposed SHIP platform, especially related to criteria of data validation and management, based on the principles described in Chapter 1.

Data management

The management of data actually includes steps of creation (collection, validation), processing, use and deletion as well.

1) Data collection

In order to create information on innovation examples and solutions, the first step is the creation of the data which can be achieved by different approaches.

As a usual default, data should be gathered by manual input of experts filling out pre-designed forms, which can be online – webforms – that is the preferred method, but also offline forms can be circulated, for example pdf x-forms, excel/access templates with built in data checks. Another option is to aggregate data automatically by software, especially if content is available in accessible way, for example in format which satisfies the criteria of open data.

Collection of information is not only a static technical task, it also has legal and logistical dimensions. The issue of authority, especially in case of the network of input providers (focal points) is very crucial one too. Input of the data is usually not a one-time accomplishment but is rather part of a regulated process where the frequency of interactions – entry and update interval, revising, removing etc. – should also be taken into account.

2) Data validation

The control of input data compliance with expected content, structure and quality is highly important, as it acts in the process flow as one of the main decision steps – to accept or reject - and has huge impact on the efficiency of next steps.

Data validation contains several consecutive steps to provide data cleaning to ensure data quality.

a) Data content. By validation of the content we mean the evaluation of the descriptive part of the innovation example or the proposed innovative solution, according to a guideline which covers the main points of criteria, perhaps also providing a scheme to indicate a certain numeric value (mark) based on the fulfillment level of a certain criteria – possibly using some weighing between different criteria – in order to not only approve or reject, but also measure the usefulness and meaningfulness of the content. Such evaluation needs manual work, performed by experts of the profession. The task of setting up the criteria for innovation and related guidance information is expected from a parallel paper contribution with this current paper.

b) Data structure. Once the data input has passed validation and evaluation of the content possibly with numeric value indicator, the next step is to check the structural correctness, which usually contains the format validation of each input field, for example the permitted syntax, value limits, length, (min,/max), and other constrains. It is possible to represent all structural requirements of an input form using XML schema definition (XSD), where each form field specifics can be described with so called simple and complex types (example see in Annex). XML schemas use the mechanism of name spaces to include other standard schemas – such as for instance metadata related ones as the Dublin Core or VOA3RAP, controlled vocabularies for subject keywords as AGROVOC, or other controlled data lists as NUTS – for different sections of the overall input structure. There are also many other taxonomies, vocabularies for keywords and categorization besides AGROVOC, like CABI, GAK (see CIARD semantic framework), and under the theme of Knowledge Organization Systems (KOS), there are many more classification schemes, thesauri, topic maps, ontologies etc. that are beyond the scope of this paper, at this stage.

3) **Process (use) of data**

This step may have different components based on the functionality of the system. Most typical ones in the context of the planned innovation platform:

a) Store and publish
- Using state of the art content management system, which fully support user administration, permissions, rights, access levels, etc
- CRUD functionality also supported on all CMS to Create-Read-Update-Delete of each CMS data record.

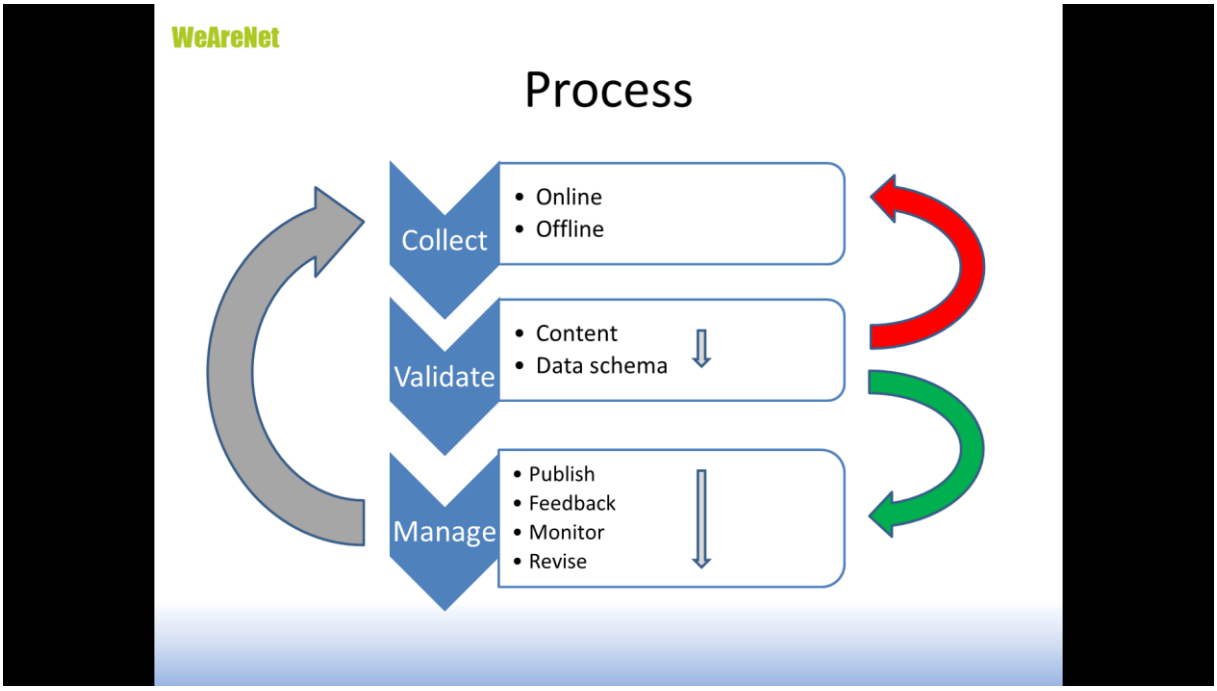
b) System monitoring
- It will be extremely important in the proposed data workflow to constantly and continuously access information about usage statistics, page ranks, link referrers, etc. Data gained from system monitoring is useful asset for improvement, evaluation and reporting.

c) Receive feedback from users
- Another valuable source of information is the feedback received directly from users of the platform, who can primarily be the end users (farmers, advisors), but also network members and experts using the workflow tools.
- The regular sources of user feedback are usually modules linked to CMS content pages, such as commenting, rating, liking, sharing, voting; and more organized ones such as guestbook, forum, survey, etc.

d) Revise the whole workflow process and operation of the system
- The revision may have impact on data collection, validation and publication methods too.

Data management also has several other usual aspects such as technical maintenance, server operation, updates, archiving, security issues, data backups, etc.

Data management workflow:



SHIP online form design for input of innovation examples

The Pre-Event Survey on Innovation Example for Small Holders and Family Farmers (Annex) was prepared and conducted before the expert consultation, which already provided some usable experience about the structure of the form, and also the response evaluation.

As usual with the preparation of surveys using questionnaire, there is always the challenge of

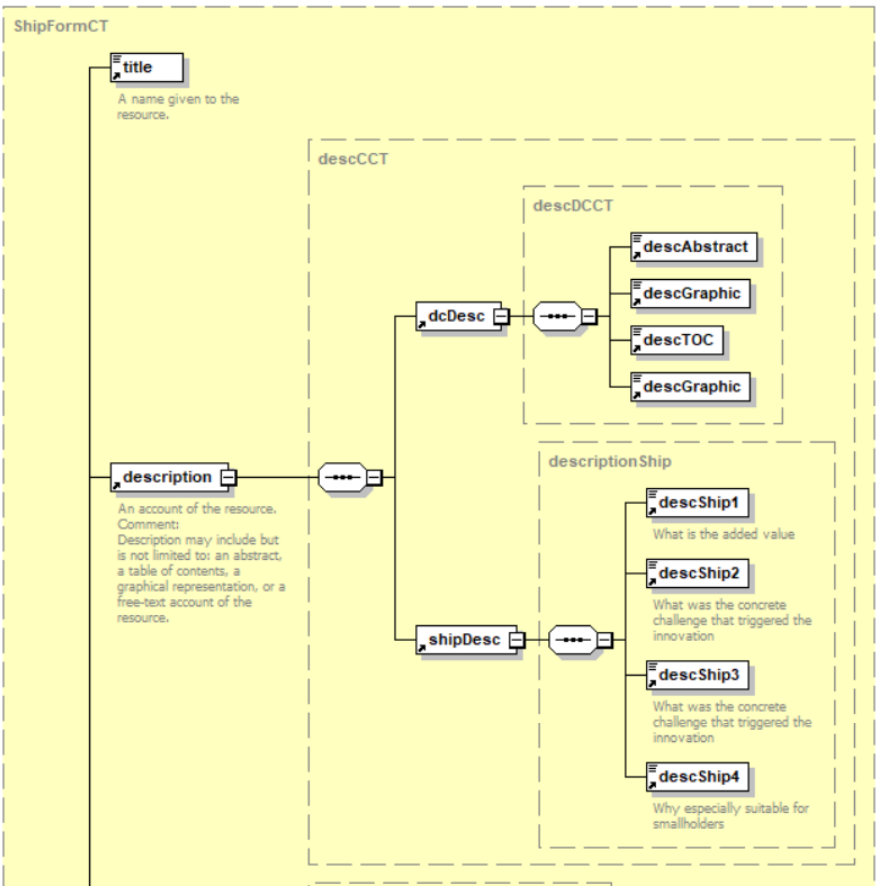
- complexity of questions; as all needed issues and data should be covered needed by next process steps
    - versus
- the easiness and quickness to fill and submit the form; to leverage the quantity of inputs.

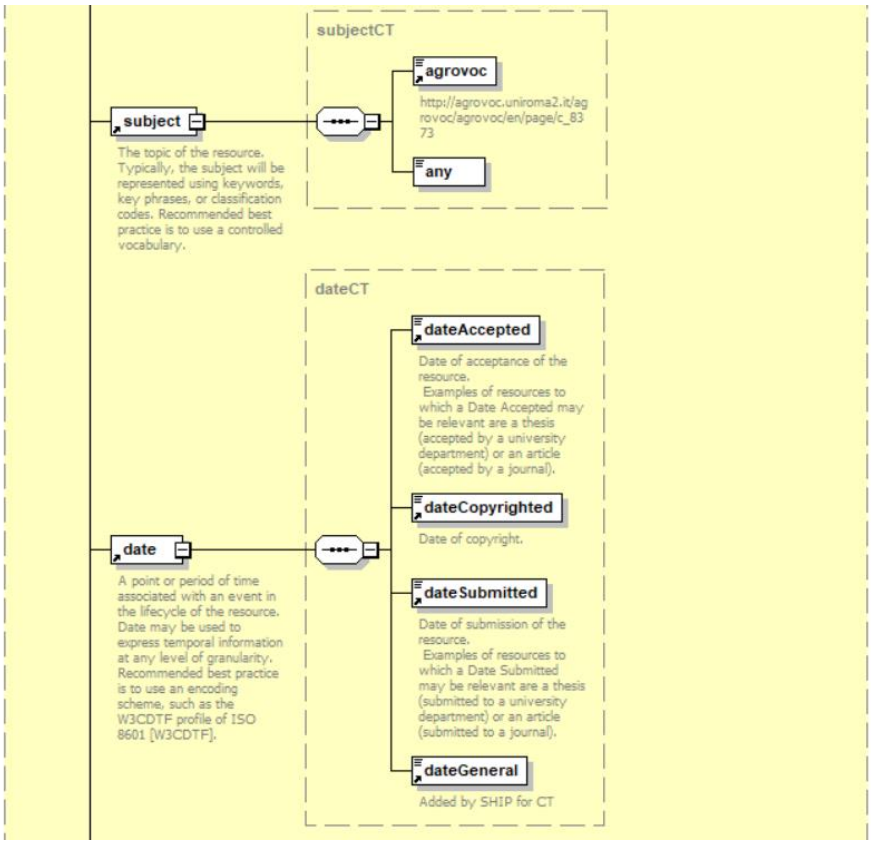
The Pre-Event Survey was made using quick and simple Google form solution, and took no respect – also because lack of such facility - to many of the criteria mentioned in this paper, for example compatibility with standard metadata scheme(s), the use of controlled vocabularies and data lists, the requirements for successful open data publishing etc.

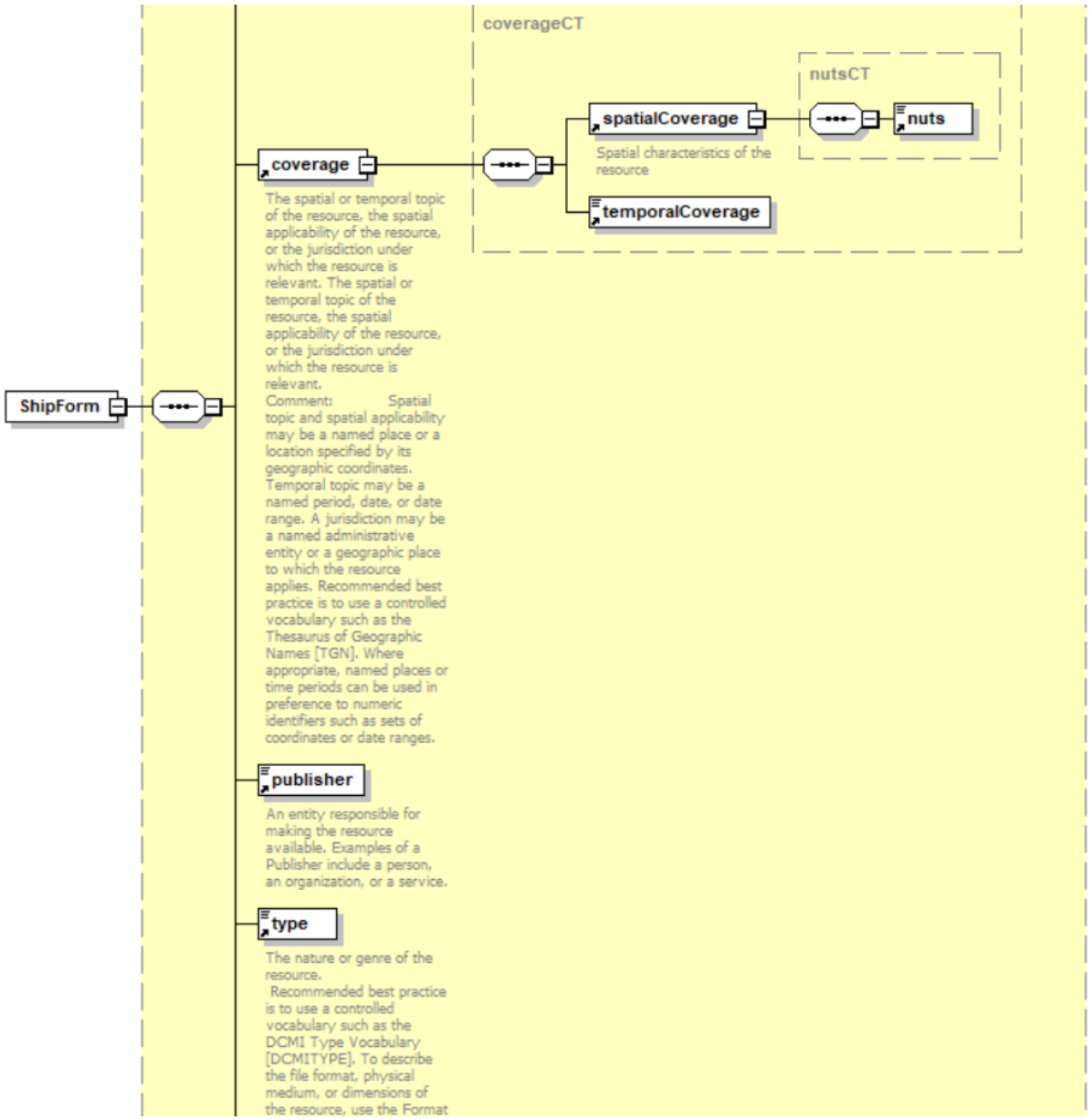
Below we provide a list of field elements that are proposed to be part of the data collection form on innovation examples.

We also demonstrate the approach by some screenshots from the designed webform, and in XML Schema (XSD) format.

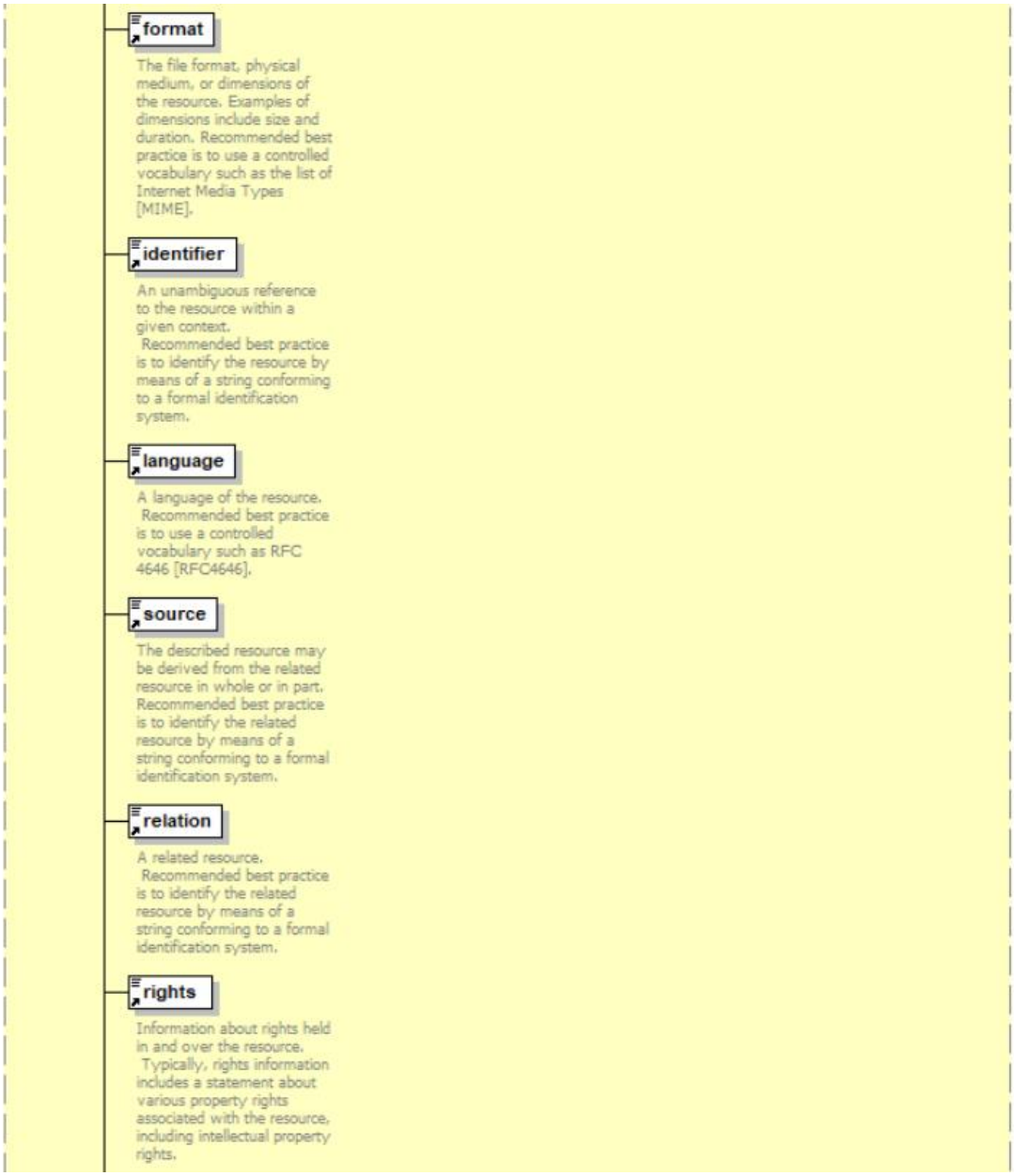
Proposed form field elements in compatibility with the Dublin Core are (example):




subjectCT
agrovoc
http://agrovoc.uniroma2.it/agrovoc/agrovoc/en/page/c_8373
any
subject
The topic of the resource. Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary.
dateCT
dateAccepted
Date of acceptance of the resource. Examples of resources to which a Date Accepted may be relevant are a thesis (accepted by a university department) or an article (accepted by a journal).
dateCopyrighted
Date of copyright.
dateSubmitted
Date of submission of the resource. Examples of resources to which a Date Submitted may be relevant are a thesis (submitted to a university department) or an article (submitted to a journal).
dateGeneral
Added by SHIP for CT
date
A point or period of time associated with an event in the lifecycle of the resource. Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF].


ShipForm
coverage
The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Comment: Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges.
coverageCT
spatialCoverage
Spatial characteristics of the resource
nutsCT
nuts
temporalCoverage
publisher
An entity responsible for making the resource available. Examples of a Publisher include a person, an organization, or a service.
type
The nature or genre of the resource.
Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format


creator
An entity primarily responsible for making the resource. Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
contributor
An entity responsible for making contributions to the resource. Examples of a Contributor include a person, an organization, or a service.



Controlled lists to be used with certain DC elements:

- Subject:
  - AGROVOC
  - Smart Farming Taxonomy
  - AgriTeach Taxonomy
  - Later: SHIP VOC
    - Generated from SHIP innovations input keywords used from AGROVOC, assisted by
      - auto-extraction of keywords (see below)
      - manual indexing, expert keyword assignment

- Spatial coverage:
    - NUTS
    - Statistical Regions For EU Candidate And ESTA Countries
    - ISO3 Country Codes

Automatic extraction of keywords

1) Keywords are often used as a short but accurate summary of documents in both physical and digital libraries. They help organize material based on content, provide thematic access, represent search results, and help navigate. Manual assignment is expensive because trained human indexers must understand the document and select appropriate descriptors according to the defined cataloging rules. The source proposes a new method that improves the automatic extraction of keywords by using semantic information about terms and phrases from a domain-specific thesaurus. The result of the developed algorithm is KEA platform.

Source: Olena (Alyona) Medelyan, KEA, http://community.nzdl.org/kea/examples1.html

2) Agrotags was developed for tagging agricultural research documents. It is a subset of AGROVOC, being much smaller with 2100, versus 40,000 terms. Agrotags was created manually by carefully examining all Agrovoc terms for their usefulness in tagging. This selected subset is further refined and validated by looking at the manually assigned keywords from the FAO AGRIS database.

3) Agrotagger acts as a module that can be used to an extension to leading repositories and advanced information management systems to automatically tag documents by controlled vocabulary such as Agrotags. User-generated tags, along with those produced by Agrotagger, would help to more effectively link agriculture-related documents to faster research and improved.

Source: T. V. Prabhakar, http://agropedia.iitk.ac.in

The software tools as results of auto tagging research can be used to accelerate the process of SHIP keyword assignment.

NUTS

The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU for the purpose of:

- the collection, development and harmonization of European regional statistics;
- socio-economic analyses of the regions;
- framing of EU regional policies (e.g. cohesion policy).

The Base-URI for the NUTS classification is: data.europa.eu/nuts/
Source: https://ec.europa.eu/eurostat/web/nuts/linked-open-data

Interoperability example

The EIP-AGRI Common Format

The EIP-AGRI common format facilitates the flow of knowledge on innovative and practice-oriented projects from the beginning to the end. The use of the EIP-AGRI format facilitates not only the exchange of knowledge, but also the contact between potential partners - farmers, consultants, researchers and all other actors - in innovation projects. It helps to build a unique pool of practical knowledge across the EU through the EIP-AGRI project database, which supports dissemination of the results of all interactive innovation projects.

Sections where information is to be provided in the Excel file:

- Instruction
- Project info
- Partners
- Keywords
- Audiovisual materials
- Website
- Practice abstract(s):

Practice abstract contains short summary for practitioners in English on the innovation project outcomes (1000-1500 characters, word count – no spaces).

- Summary should at least contain the following information:
    - Main results/outcomes of the activity (expected or final)
    - The main practical recommendation(s): what would be the main added value/benefit/opportunities to the end-user if the generated knowledge is implemented? How can the practitioner make use of the results?
    - should be as interesting as possible for farmers / end-users,
    - using a direct and easy understandable language
    - pointing out entrepreneurial elements which are particularly relevant for practitioners (e.g. related to cost, productivity etc).
    - Research oriented aspects which do not help the understanding of the practice itself should be avoided.

Source: https://ec.europa.eu/sfc/en/community/document/template-eip

SHIP's relation to EIP format

We intend to take into attention the EIP format, especially its approach for practical use (see above) and for possibilities in interoperability, however, we plan to:

1) go more structured - the EIP abstract field is one text field, while we will use several descriptive sections and options
2) focus on innovation for smallholders
3) collect those that have been adding value/validated by practice.

Interoperability interface

The SHIP platform will enable the opportunity to generate data on innovative solutions in Open Data format. In the practice it will mean the mapping of SHIP data fields to metadata which can be harvested by EIP-AGRI (if implemented) or in simple file format (CSV or

XML) that contains the data in a structure that is easy to download, understand, process and integrate into their system, if required.

Project design example

The meeting "Regional Expert Consultation on Knowledge Sharing For Agricultural Innovations Applicable For Smallholders" was organized in Godollo, Hungary by FAO, WeAreNet and GAK St Istvan University, participated by specialists from EIP-AGRI (EU Commission) and from 26 nations to follow up with the regional conference, and proposed the setting up of a network and platforms which is to collect good examples on innovations for smallholders.

For the realization of the targeted platform goals, beyond the follow up activities of the regional expert consultation that provided guides and recommendations for developing the innovation platform and initial data input on voluntary base, we designed a concrete project proposal that goes beyond in the next steps:

1. identifying all possible sources of innovation examples and collecting the most relevant ones
2. collecting data both on the solution side (supply) and the problem side (farmers' demand, challenges)
3. pairing and connecting problems with innovations for the solution (match-making approach).

During the regional consultation in Godollo, experts discussed about networks and platforms in the region which aim to improve smallholder performance and livelihood by new methods and innovative solutions. It was a common observation that while there are many providers visible from the technology side and a lot of information is available from science, research and extension, we know very little about the end-user side, i.e. about concrete problems smallholders face during their daily work, and how these relate to the offered solutions. Another observation stated that suggested innovations should be capable to directly and specifically respond to the needs coming from the field. Therefore data about practical problems and questions should also be collected and matched against - or paired with - the offered innovations. The objective of the current project proposal is to set up a dialogue platform and a pilot action to develop smallholder and family farm support mechanism model to be used in V4SEE and other relevant geographic regions facing similar challenges by the creation of:

- a network of data collection from the field - advisers, extensionists, field day demonstration experts, on farm researchers,
- a network of innovation producers - researchers, scientists, private companies,
- a method to collect, evaluate and publish data from both the practitioner (smallholder) and the provider side.

The main purpose is to collect and classify smallholder and family farms' problems by typology, create and track each issue on the information system (platform), experts matching, pairing, connecting innovations with problems for solutions, the results channeling back to the end users and also storing in the advisory knowledge base.

1. Two advisors in each country regularly working with smallholders will use a template – or field data collect app (e.g. EpiCollect5) – to register problems and needs of end

users in a structured way – using problem taxonomy, and then upload it to the online platform (at least 40 farmers per country).
2. Two innovation providers from each country – such as researcher, developer, knowledge broker, etc. - will collect and upload innovation examples suitable for smallholders and family farms, by similar taxonomy structure.
3. Selected experts will match – pair and connect – problems with innovations for solutions. This can be made at the top level (full match), or sub-level (sub-problem / sub-solution, partial match).
4. The problem-innovation-solution triplets will be stored and made searchable, using open data and semantic web technology.
5. The concrete solution for the actual problem issue will be channeled back to the specific end user too.
6. The whole process will be monitored, analyzed and fed back for improvement and policy information.

Reason to address the issue on regional level:

By a recent publication from FAO, the state of innovation ecosystem varies by country to country in Europe and Central Asia, including the V4 + WB region, being fragmented within the countries as well. A strategic approach is needed to take advantage of the potential of innovation for small holders and family farms that is the dominant way of pursuing entrepreneurship in rural areas, especially in remote parts of South East Europe, but also significant in the V 4 countries. Small enterprises cannot maintain a separate unit to carry out advanced management tasks, therefore they depend very much on the individuals to use a new method and on the support mechanisms that identify the problem and can provide guidance on the solution, taken into attention the specifics of regional and local aspects.

Possible risks for the project success (data collection and use):

1. Communication-conservative mentality of smallholders and family farms that may have difficulties to open up with visiting field experts and to provide information on their problems and needs. To manage this risk, agricultural advisers should be involved who are already known and trusted by the visited entrepreneurs.
2. Need of adapting the offered innovative solutions. Available information on suitable innovation examples must be filtered not only related to size, sectors and specific need, but also regional aspects, including environmental factors, social attributes, etc., a complex knowledge to be possessed by selected experts on the innovation provider / research / academia side.
3. Reaching the targeted indicators. Experts will be selected according their well stablished position in their fields, both on the problem and solution side. Data from the field will be selected during the season when practitioners are actually most like to face problems, i.e. during spring and summer months.
4. Lack of smallholders' willingness to accept the proposed innovative solution. It is necessary to use practical demonstration methods, to not only provide description and guidance in document format (paper/electronic), but more interactive and hands-on techniques, videos, in the field presentations, field days, testimonies from other smallholders, etc.

Activities to share the results of the project outside the partner organizations:

Results will be shared during numerous national and international actions that each partner is conducting in many different areas**.** At the international level, partners are quite active in EU and UN FAO projects and events, at least with annual frequency. Experts, who are to be assigned for tasks of visiting smallholders to collect data on problems and needs, are advisers working in national networks and shall share results with many other practitioners. Main related projects currently: AgriTeach 4.0 Erasmus+, BalkanMed Innova Erasmus+, H2020 BOND, SKIN.

Activities in the future, which will build on the results of this project:

Based on expected good experience and the tools developed, they plan to introduce project results to all other Eastern Partnership and Western Balkan countries using the acquired process know-how. This will be facilitated by Small Holder Innovation Platform, maintained by WEARENET and associated experts and members of the network. Besides the international dimension, the network can be expanded at the country level as well, to allow any agricultural adviser, field expert, business consultant and innovation broker etc. to register and use the platform as a tool for better serving their small holder clients

Deliverables:

1) Match-making platform

An integrated information system will be put in place for collecting, managing and publishing data, issue tracking problems, monitor the process and report on the results. The platform will be based on open source CMS, augmented with proprietary system extensions for specific functionalities to the core platform. The added module records, tracks and reports issues and their related interactions between small holders and experts. It is the center of the system, assisting with the full life cycle of the process - from the point of first contact between the end user and adviser - through the monitoring and evaluation of the service. The platform will be ready to receive inputs before the start of field data collection, including the administration of taking records of the visits (logbook signed by small holder and visiting expert). The match making and reporting module will be fine-tuned during the interval of data collection and pairing process.
Direct target groups:

- Field experts on small holder problems: 10. Selection/outreach: Internal
- Research, academia on innovation examples: 10. Selection/outreach: Internal
- Other actors, small holders, advisers, researchers, innovation providers etc who wish to sign up: 20
- Selection/outreach: Online, registration controlled and data entry moderated, quality checked.

Dissemination/promotion: Website will be promoted on all printed materials, including internal ones (templates, reports) and external, public ones (booklet, presentations etc). It will also be promoted online on partners' web pages and via other networks of the applicant.

2) Data collection and matching

Two advisers in each participating country working in the field will collect issues about problems and needs from 40 smallholders - at least two problems per farmer to be entered into

the issue tracking system - which could possibly be solved by the use of innovative methods. Two researchers in each country will collect at least 20 examples of possible innovative solutions that are expected to be applicable to the needs of small holders. Altogether 200 smallholders will be reached by the project, 400 problem issues and 100 examples for innovative examples will be collected and uploaded to the "match-making" platform, where the experts' role will be to analyze the issues, connect problems and solution at the top level (if possible), or if needed at sub-problem to sub-solution level(s). Matched issues will be channeled back to the small holders, by experts revisiting them (at no cost) and presenting the solution, which will have to be approved or rejected by the end user, signing the project template and validating the whole process. This will also be a key indicator figure for the final project report.

Innovations can be type of technology, process, social, organizational and environmental. It should be taken into attention the suitability for the small-scale holders, the internal and external factors for adoption of the innovation, and the economic viability of the desired innovation practice. Recommended innovation practice should be relevant, achievable, feasible, replicable, sustainable and creditable, and according to market needs.
Direct target groups:

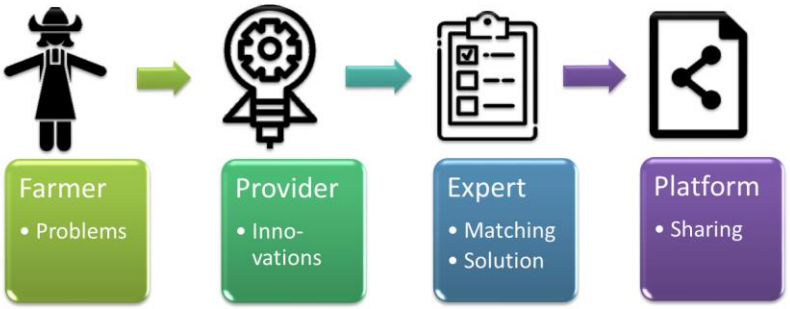
- Innovation providers: 100. Selection/outreach: Desk research, journals, periodicals, visiting events, exhibitions.
- Smallholders: 200. Selection/outreach: According to the EU methodology, a small holder farm is defined by size and output. Field experts, advisers will visit selected entrepreneurs in their own environment.

Dissemination/promotion: The action will be promoted widely on the websites of each partner so that any smallholder can voluntarily join the process by his *I* her own adviser or even enter issue record on the system alone (it will be quality assured at the system level). Same applies to the academia and the innovation provider sector, which will also be publicized online and made open for entries and then quality checked.

Role of the applicant and project partners:

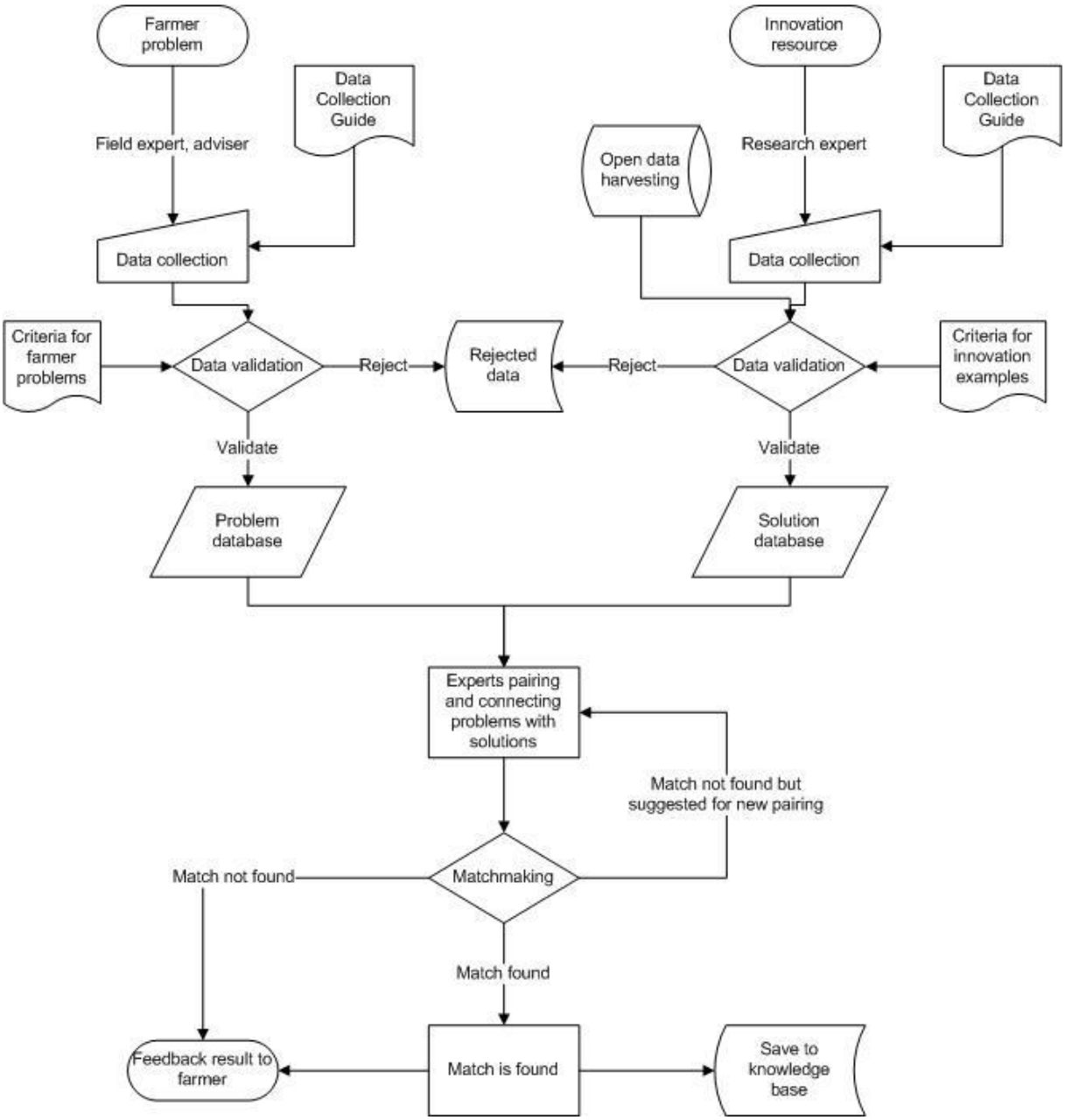
Each partner and the applicant will collect data from the field and from resources of innovative solutions (articles from scientific literature, description of good examples, service providers, etc). The applicant will monitor the process, coordinate the matching expert work, generate reports and maintain the system. Partners will feed back solutions and validate the process, by requiring small holder's approval or rejection, enforced by signature.

Problem solving process

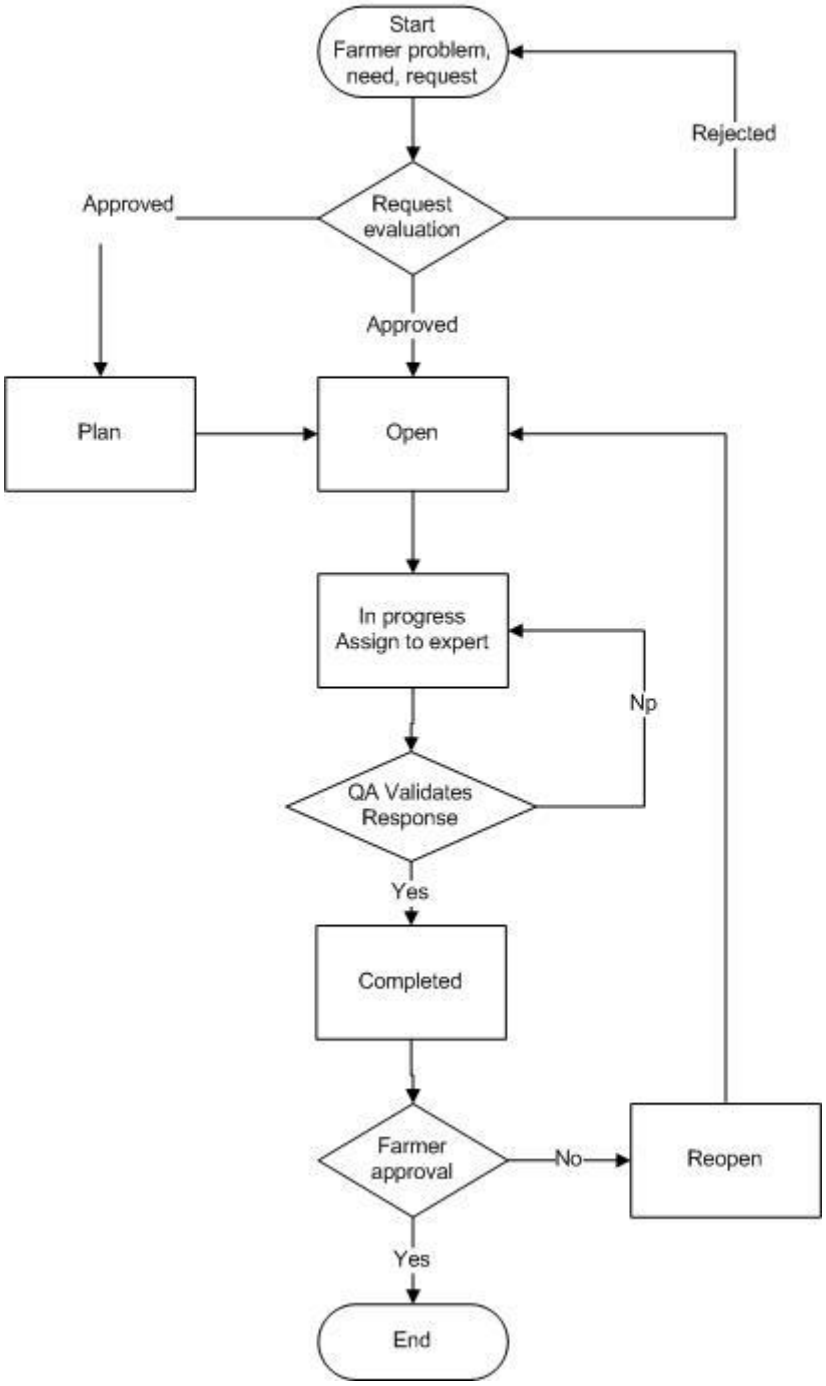


WeAreNet

SHIP matchmaking process



SHIP Issue Tracking

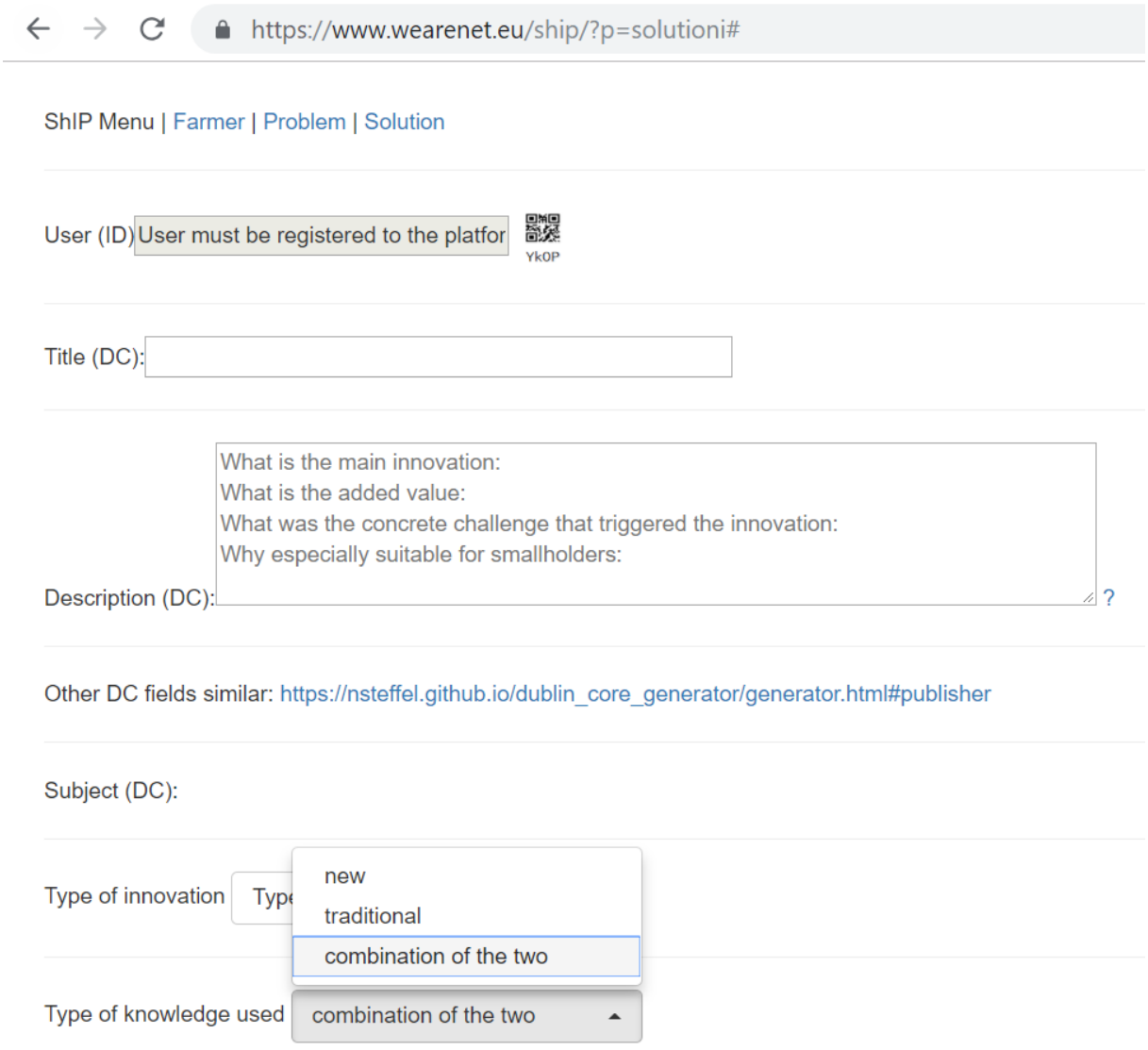


Online form components example

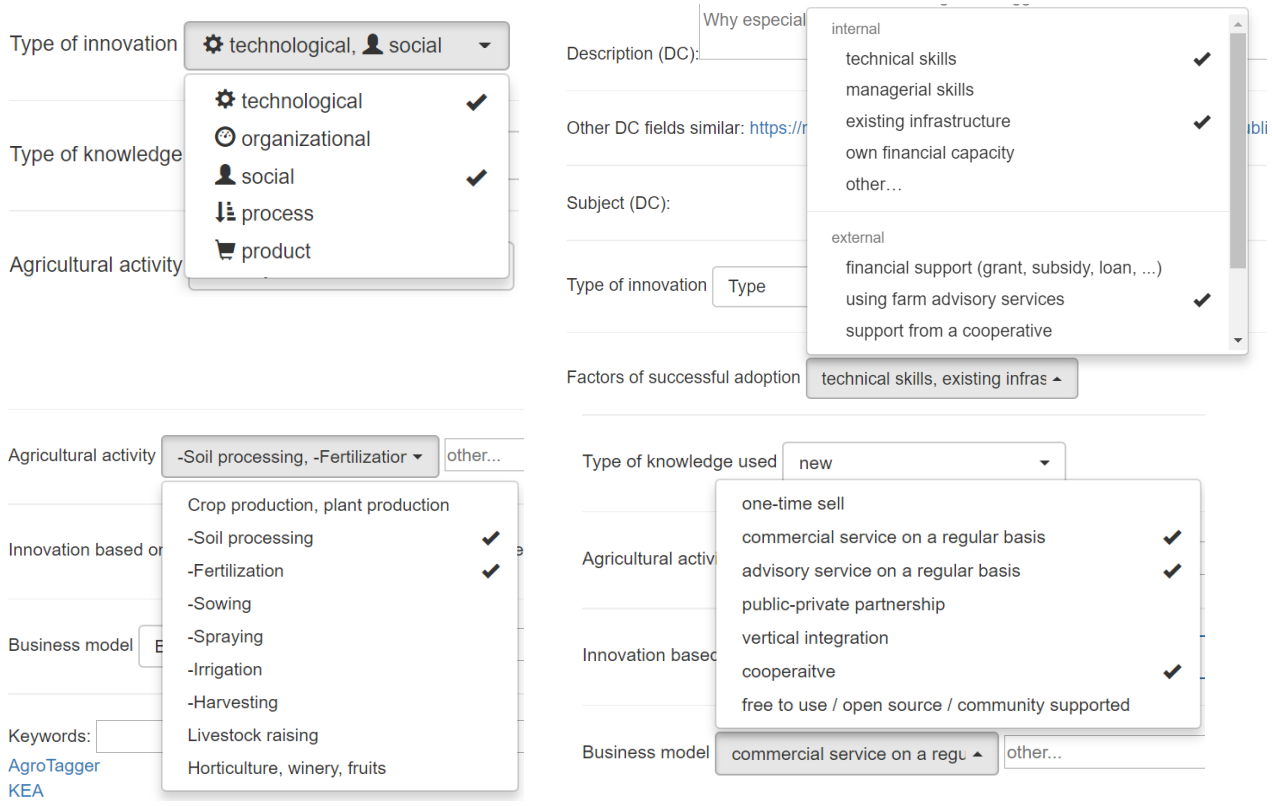
We recommend to develop professional web form, using:

- user friendly graphical user interface (GUI) components for form input elements, selects, checkboxes, radio buttons, etc.
- responsive layout for compatibility with mobile devices
- custom validations

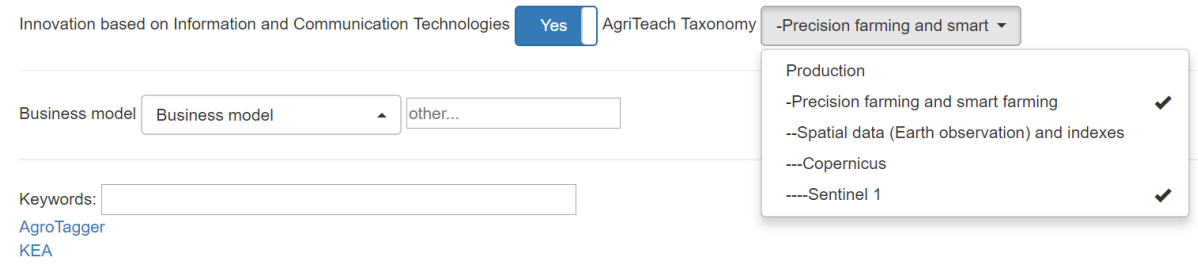
Only registered users can access the form, the application will generates a unique ID for the session that can be used later as reference code for accessing saved data for further work.



Multiple select boxes with icons and with option groups:



Toggle replacing radio button:



The above examples belong to the first item of the 4 main functionalities planned for the SHIP platform (initial approach):

1. Innovation input
2. Farmer typology (with client database)
3. Farmer request, issue tracking system
4. Matching (connecting) innovations with problems for solution

References

Using Metadata Description for Agriculture and Aquaculture Papers. P. Šimek, J. Vaněk, V. Očenášek, M. Stočes, T. Vogeltanzová

Introducing a content integration process for a federation of agricultural institutional repositories. Vassilios Protonotarios1,4, Laura Gavrilut1,4, Ioannis N. Athanasiadis1,2, Ilias Hatzakis1, Miguel-Angel Sicilia3

Agrotags – A Tagging Scheme for Agricultural Digital Objects. Venkataraman Balaji, Meeta Bagga Bhatia, Rishi Kumar, Lavanya Kiran Neelam, Sabitha Panja, Tadinada Vankata Prabhakar, Rahul Samaddar, Bharati Soogareddy, Asil Gerard Sylvester, Vimlesh Yadav

Text mining in agriculture: The AgroTagger keyword extractor. AgroTagger classifier. Leveraging the knowledge encoded in AGROVOC for AGRIS items. KARWOWSKI, W., WRZECIONO, P., Methods of Automatic Topic Mining in Publications in Agriculture Domain, Information Systems in Management (2017) Vol. 6 (3) 192−202.

AGROTAGS: a contribution towards improved digital information management in agricultural research, 2010. T.V. Prabhakar, Lavanya Kiran Nelam, V.Balaji, in Annals of Library and Information Studies

Domain Independent Automatic Keyphrase Indexing with Small Training Sets. Olena Medelyan and Ian H. Witten Department of Computer Science, The University of Waikato, Hamilton, New Zealand

Thesaurus-Based Index Term Extraction for Agricultural Documents. Olena Medelyan, Ian H. Wittenbm Department of Computer Science, The University of Waikato, Hamilton, New Zealand

Human-competitive automatic topic indexing. Olena Medelyan - Thesis

Automatic keyphrase extraction from scientific articles. Su Nam Kim, Olena Medelyan

Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vollume 3, Issue 4, pp. 257-279. O. Medelyan, I.H. Witten, A. Divoli, J. Broekstra. 2013.

Establishing basic infrastructural and methodological background of Semantic Web in the domain of food and agriculture in Hungary. Laszlo Gabor Papocsi and Ferenc Kulcsar GAK SZIE - University of Gödöllő

Smart Farming Technologies – Description, Taxonomy and Economic Impact, Athanasios T. Balafoutis, Bert BeckSpyros Fountas, Zisis Tsiropoulos, Jürgen Vangeyte, Tamme van der WalI. Soto-Embodas, Manuel Gómez-Barbero, Søren Marcus Pedersen, 16 November 2017

Using of Automatic Metadata Providing, P. Šimek, M. Stočes, J. Vaněk, J. Jarolímek, J. Masner, I. Hrbek, 2013

Annex

Summary of pre-event data collection responses about innovation examples for smallholders and family farmers

Food and Agriculture Organization of the United Nations

WeAreNet

Overview of pre-event survey responses on examples of innovation for smallholders and family farms

Laszlo Gabor Papocsi
GAK / WeAreNet

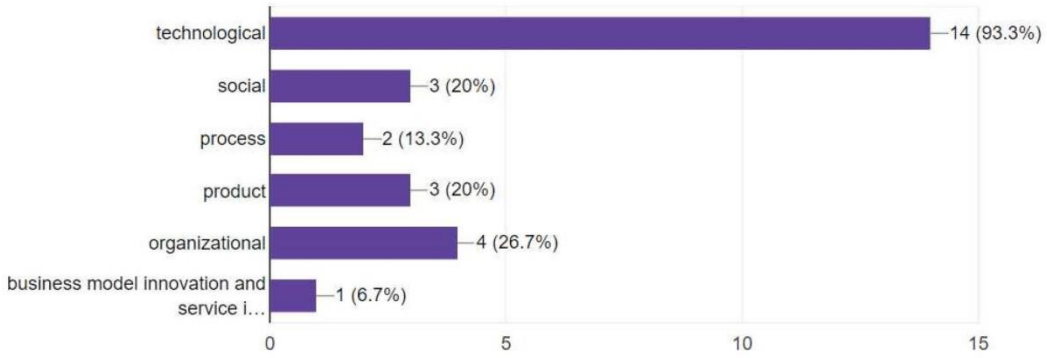
FAO Expert consultation on knowledge sharing for agricultural innovations applicable for smallholders and family farmers in Europe and Central Asia

Godollo - Hungary, 10/09/2018 - 13/09/2018

WeAreNet

Type of innovation:

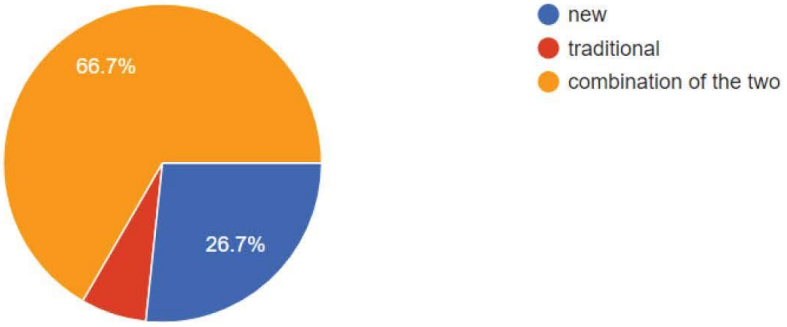
15 responses



WeAreNet

Type of knowledge used:

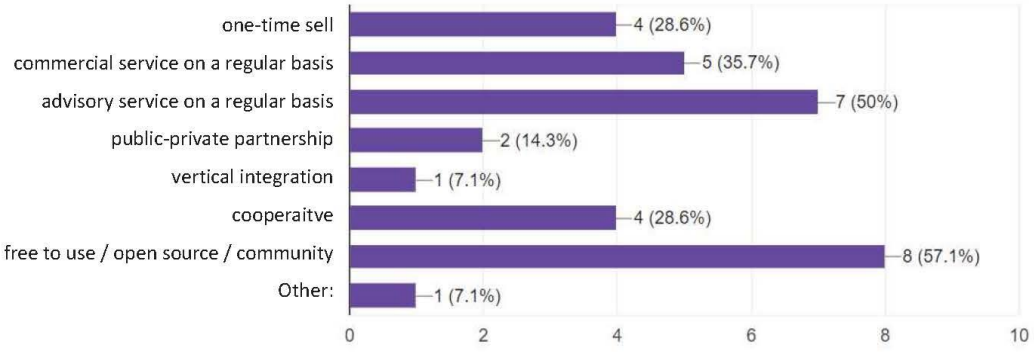
15 responses



WeAreNet

Business model

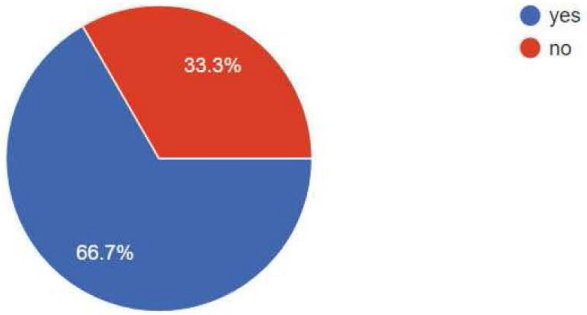
14 responses



WeAreNet

Innovation based on Information and Communication Technologies

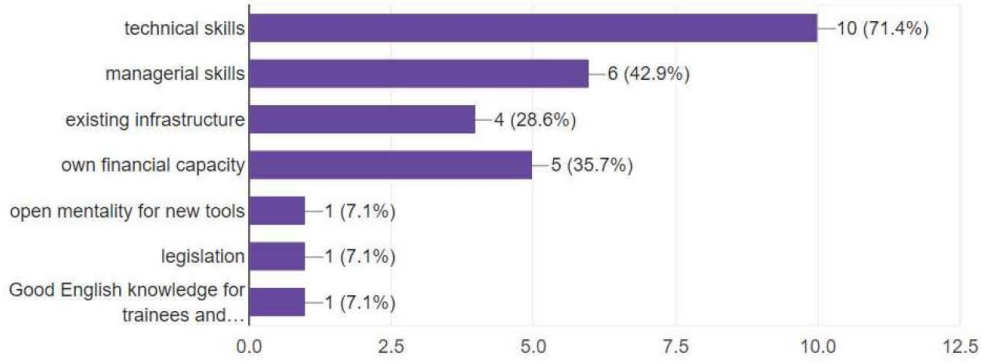
15 responses



WeAreNet

Factors of successful adoption - internal

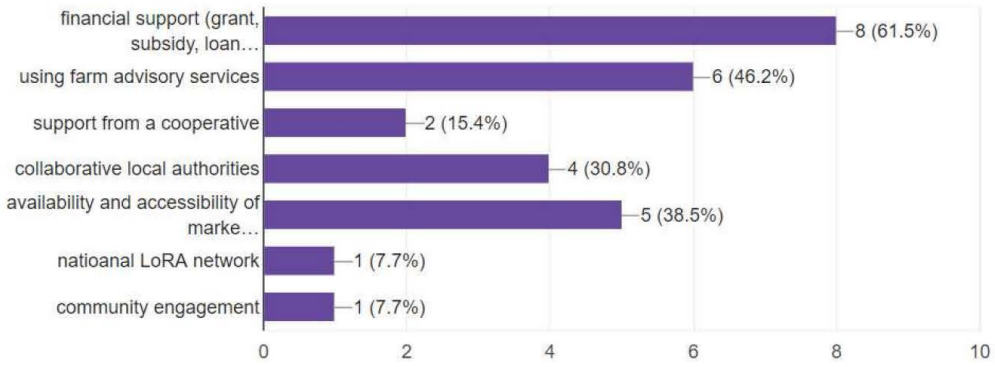
14 responses



WeAreNet

Factors of successful adoption - external

13 responses



WeAreNet

Keywords (tags) (sector, type of technology etc)

12 responses

pest forecast, smart phone
Organic crops, wind power
planting
LORA, IOT
family farms, smallholder agriculture, small-scale farming, farm tourism, rural tourism, traditional agriculture, sustainability, promotion, marketing, web applications
drone, maps, crop health, data analysis
short food supply chain, cooperation, local food supplieres
beekeeping, ICT, remote sensing, precision farming, small farms and agricultural households, family farming
advisory service, organic farming, small-scale farmers, continuous advice and education for farmers, on-farm agricultural research, on-farm education, Green Training Center
Web School, virtual classroom, AGRI, startups,
Bar-coding System, early warning system, drone using, face recognition system, radio frequency controlled

What is the innovation

1. Affordable remote monitoring of insect traps
2. Precise agriculture
3. Measures for adaptation to climate change
4. Wind powered water pump for an organic farm
5. Hand planter that includes 2 seed drums, and 1 fertilizer drum.
6. LoRA WAN private data communication network
7. Integrated search platform which offers a possibility to promote such family farms in Slovenia which provide farm holidays
8. Web platform that converts drone captured images of farm fields into crop health data.
9. The new social interactions between the producers and local community and the new way of selling the products.
10. Remote honey production control
11. Green Training Center (GTC) in Dzoraghbyur village, Armenia
12. Knowledge sharing for the use of renewable energy resources for further processing/preparing (drying) of non-wood forest products
13. Virtual based tools and novel approach for learning and vocational training.
14. Digital usage in aggregated agricultural land:
15. LandVoc is a controlled vocabulary covering any concepts related to land governance.

Challenge

- Small farmers cannot monitor numerous traps daily / weekly if their plots are scattered around the municipality / region
- Knowledge of farmers; Infrastructure; Finances
- To bring water from an irrigation system channel up to a higher elevation field without the use of costly and polluting technology such as diesel pumps.
- Manual work too tiring, waste of inputs, inconsistent yield
- Identifying and montoring cattles in the pastoral type of a grazing system, in remote places on the pastures of the Hungarian Puszta
- Analysing drone captured images
- Problems that smallholders had with the wholesalers.
- Time and classic trail/error approach resulting in many hours and traveling costs for checking the beehives
- Collection of big amounts of products with no profit
- Prepare the training curricula that fits the needs and priority of target groups in different countries of the region, in order to attract as many users as possible
- Low income and especially young farmers were tend to migrate to the urban cities.
- As web aggregator, the Land Portal needs to map content from several sources into a common model.

Why for smallholders

- Cheap, advisory assisted, easy to learn to use, can be applied based on open source guidance
- More efficient and more productive
- Has low to no operating costs, easily applied to most terrains and applicable to remote regions with poor infrastructure.
- As "green seeder" part of the small farm toolbox, see URL
- No monthly cost if own network is used, one network can serve a list of useful functions in the farm by bi-directional active low power sensors
- Farm-tourism is an additional source of income for family farms.
- Free/cheap and decrease knowledge barriers
- Smallholders are participants of short food supply increasing their farm income.
- Low cost solution of (almost) precision farming in beekeeping
- Show smallholders an effective business model for commercial farming and help them adopt effective and efficient farming and business practices.
- Create an opportunity for income generation activity
- Training curricula will aim especially small holders, SMEs, young entrepreneurs
- Digital agriculture is also an opportunity to solve the structural problems of agriculture in Turkey.