



LESSONS ON CREATING A TRULY MULTI-LINGUAL PORTAL



DELIVERABLE

Project Acronym: **Organic.Lingua**

Grant Agreement number: **270999**

Project Title: **Organic.Lingua: Demonstrating the potential of a multilingual Web portal for Sustainable Agricultural & Environmental Education**

D8.7 Organic Lingua White Paper

Revision: Final

Authors:

Benjamin Cave (BCU)

Vassilis Protonotarios (UAH)

Marc Dymetman (XER)

Alessio Bosco (CELI)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Revision history:

Revision	Date	Author	Organization	Description
0.1	21/02/2014	B. Cave	BCU	First draft ToC and Chapter 1
0.2	27/02/2014	N. Marianos, G. Stoitis	AK	Additional Content
0.3	28/02/2014	B.Cave	BCU	First Complete Draft
0.4	06/03/2014	B. Cave	BCU	Second Complete Draft
0.5	08/03/2014	A.Bosco	CELI	Input to Chapter 3
0.6	10/03/2014	M. Dymetman	XER	Input to Chapter 4
0.7	12/03/2014	V. Protonotarios	UAH	First Review and Input to Chapter 5
0.8	18/03/2014	B. Cave	BCU	Final Draft
0.9	18/03/2014	S. Ruston	BCU	Review & QA
1.0	18/03/2014	B. Cave	BCU	Final

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of contents

TABLE OF CONTENTS.....	3
TABLE OF FIGURES	4
EXECUTIVE SUMMARY	5
1 INTRODUCTION	7
1.1 SCOPE.....	7
1.2 AUDIENCE	7
1.3 DEFINITIONS	7
1.4 STRUCTURE.....	8
2 ORGANIC.EDUNET: BEFORE AND AFTER ORGANIC.LINGUA	9
2.1 BEFORE ORGANIC.LINGUA.....	9
2.2 AFTER LINGUA.....	10
2.3 REVAMPING THE ORGANIC.EDUNET WEB PORTAL	12
2.3.1 <i>Multilingual content discovery</i>	12
2.3.2 <i>User evaluation of automatic translations and improvement</i>	13
2.3.3 <i>User generated content</i>	14
2.3.4 <i>Domain-powered content discovery</i>	15
3 ENABLING MULTI-LINGUAL CONTENT SEARCH – LESSONS FOR THE FUTURE	17
3.1 OVERVIEW OF ORGANIC.LINGUA CLIR SOLUTION	17
3.2 LESSONS LEARNED – CONTENT SEARCH	18
3.2.1 <i>Plan the proper degree of support for the selected languages.</i>	18
3.2.2 <i>Use High Level Interfaces to Wrap Language Resources.</i>	18
3.2.3 <i>Reuse Public Knowledge</i>	19
3.2.4 <i>Link the Portal Knowledge to the Public Knowledge</i>	19
3.2.5 <i>Involve Users</i>	19
3.2.6 <i>Caching Mechanism</i>	20
4 ENABLING MACHINE TRANSLATION- LESSONS LEARNED	21
4.1 OVERVIEW OF ORGANIC.LINGUA MT SOLUTION	21
4.2 LESSONS LEARNED – MACHINE TRANSLATION	22
4.2.1 <i>Plan the proper degree of support for the selected languages</i>	22
4.2.2 <i>Provide careful estimates of the effort required for in-domain adaptation</i>	22
4.2.3 <i>Identify the test sets carefully at the beginning of the project</i>	22
4.2.4 <i>Anticipate risks and have fall-back plans</i>	22
4.2.5 <i>Use High Level Interfaces to Wrap Language Resources.</i>	23

4.2.6	<i>Caching Mechanism</i>	23
5	ENRICHING YOUR METADATA – LESSONS FOR THE FUTURE	25
5.1	OVERVIEW OF THE ORGANIC.LINGUA METADATA & UGC SYSTEM	25
5.2	LESSONS LEARNED – METADATA ENRICHMENT	26
5.2.1	<i>Even automatic metadata harvesting/ingestion can be problematic</i>	26
5.2.2	<i>The use of the appropriate metadata authoring tool can help metadata authoring</i> ..	27
5.2.3	<i>Multilinguality matters</i>	27
5.2.4	<i>High quality metadata need time and guidance</i>	27
5.2.5	<i>Use the power of the community</i>	28
5.3	CONNECTION WITH GLN	28
6	ENGAGING YOUR USER BASE – LESSONS FOR THE FUTURE	29
6.1	LESSONS LEARNED – USER BASE	29
6.1.1	<i>Get Influencers Onside Early</i>	29
6.1.2	<i>Identify the target audience</i>	29
6.1.3	<i>Keep the stakeholders engaged</i>	30
6.1.4	<i>Tie-In Events Provide Real Hands-On Feedback</i>	30
7	HOW TO USE THE ORGANIC.EDUNET SOLUTION	31
7.1	INITIAL STEPS.....	31
7.2	OPTIONS FOR PUBLISHING CONTENT THROUGH THE ORGANIC.EDUNET WEB PORTAL	31
8	CONCLUSION	33
	ANNEX: THE ORGANIC EPRINTS DISCOVERY SPACE	35

Table of Figures

Figure 2.1. The UGC widget in the Organic.Edunet Web portal	11
Figure 2.2: Overview of the multilinguality of the Organic.Edunet resources	13
Figure 2.3: User-provided metadata translation	14
Figure 2.4: The basic steps of the User Generated Content workflow	14
Figure 2.5: Domain-specific content retrieval in Organic.Edunet	15
Figure 5.1: Some of the available search filters based on educational context (a), resource language (b) and resource format (c)	25

Executive Summary

The Organic.Lingua project set out with a simple mandate: To transform a successful and vibrant education portal into a truly multilingual end-user experience that facilitates easier access to information and content. Despite the deceptive simplicity of this overarching goal, the project found our resulting work to be both challenging and informative.

The entire Organic.Lingua team felt that it would be beneficial for those working in the area of multi-lingual content in the future to have a reference document which sets forth the best practices we have developed and give a short overview of the challenges we faced. To this end, our white paper has been structured to provide useful recommendations on the different aspects of developing a truly multilingual portal.

The paper opens with details of the Organic.Edunet portal which formed the focal point for Lingua's work. It shows how with the addition of targeted, domain-adapted translation capabilities, a collection of resources, no matter their original language, can be rendered accessible and informative to people across Europe and beyond. The following sections provide further details of the developments necessary to realise this goal. First the paper examines the role of multi-lingual search functionality in giving easy access to resources. Second, the paper examines the role of clean, meaningful metadata in improving resource discoverability. Finally, the paper suggests some qualitative recommendations for engaging platform users and effectively communicating the benefits of multi-lingual search. The paper concludes with two practical sections designed to help those wishing to repurpose or sustain Organic.Lingua's tools in their own work. First, an instructional section provides readers with an overview of the Organic.Edunet solution as it currently exists. Finally, the paper concludes with practical steps for embedding the functionalities of Organic.Lingua in your own work.



1 Introduction

1.1 Scope

This deliverable is the final White Paper from the Organic.Lingua project. It provides information about the advantages of multilinguality options provided through the Organic.Edunet Web portal and other tools by presenting the Organic.Lingua experience. It covers aspects like the machine translation tools used for providing domain-specific translations of metadata, the importance of the multilingual search for retrieving learning resources, multilingual metadata annotation for facilitating the indexing and retrieval of learning resources and issues related to the engagement of users in a learning portal.

This paper describes the experience gained from the Organic.Lingua project so lessons learned can be reused by various types of stakeholders, such as related projects and initiatives, organizations wishing to make use of a multilingual portal, metadata annotators who face issues with the multilinguality aspects of their collections, and repository managers who wish to provide multilinguality to their resources etc.

1.2 Audience

This document addresses a wide variety of stakeholders from project managers working with multilinguality aspects in their projects, repository managers and metadata annotators, web portal managers as well as anyone working with multilinguality issues in the context of their work.

1.3 Definitions

AgLR: Agricultural Learning Repository Tool, a tool that facilitates the management of agricultural educational resources and collections in the context of the Organic.Lingua project.

AP: Metadata Application Profile, a set of metadata elements taken from a specific metadata standard/schema in order to meet specific requirements in a specific context.

IEEE LOM: Standard for Learning Object Metadata published by the Institute of Electrical and Electronics Engineers Standards Association, New York.

OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting

UGC: User Generated Content



1.4 Structure

Chapter 1: contains an overview of this document, providing its Scope, Audience, and Structure.

Chapter 2: provides an overview of the Organic.Edunet Network

Chapter 3: provides a number of benefits of connecting to the Organic.Edunet Network for content providers

Chapter 4: provides practical information for connecting to the Organic.Edunet Network

Chapter 5: defines the contribution of this deliverable to the Organic.Lingua vision

2 Organic.Edunet: Before and After Organic.Lingua

Organic.Edunet has been the focal point of the Organic.Lingua project, around which our multiliguality work has been based. Therefore, it is appropriate to begin with a brief examination of the Organic.Edunet Web portal and highlight areas that have been improved by the Organic.Lingua solution.

2.1 *Before Organic.Lingua*

Organic.Edunet represents a network of resources drawn from content providers with educational materials in organic agriculture, agroecology and other green topics like ecology, sustainability, energy, biodiversity and environment. The purpose of creating Organic.Edunet was to provide a single point-of-entry for discovery of resources often held in small collections which end-users may be unaware of. By centralising all these resources into a single, trusted portal (www.organic-edunet.eu), Organic.Edunet helped to create a community of users who in turn helped to further strengthen and develop the resources available. The Organic.Edunet network was initially formed in the context of the Organic.Edunet eContentPlus project, which started in 2007 and ended in 2010. The initial network consisted of eleven (11) collections provided by a variety of content providers, including schools, universities, associations, educational institutions etc., as well as two user communities. The project provided a critical mass of metadata records, a number of which are available in two or more languages.

The resources available through Organic.Edunet cover all levels of formal education in addition to lifelong learning and vocational training. One major challenge when considering the implementation of multi-lingual search is the diversity of content types available through the portal. Everything from lesson plans through to multimedia files and academic articles can be discovered through Organic.Edunet.

Prior to the integration of Organic.Lingua functionality, Organic.Edunet supported four types of user search functions: text-based, tag-based, browse using predefined filters and navigational search. In addition, to support the core user demographic of educationalists, competency-based search allowed customisation based on the intended audience level. Navigational search was also available based on the classification of the resources according to the Organic.Edunet ontology and allows users to browse through concepts in a user-friendly tree-like interactive diagram.

Organic.Edunet aggregates metadata records, so the majority of the interconnected repositories and collections support the exposure of metadata through an OAI-PMH target. Organic.Edunet has its own IEEE LOM-based metadata application profile, so a transformation/mapping between the metadata elements used in a collection and the Organic.Edunet ones may be needed.



In summary, Organic.Edunet is a fully-capable educational portal developed expressly to serve the needs of the agricultural education community. In every aspect of this portal, the user experience was customised to support a tailored environment for connecting people with the right resources. In addition to the work Organic.Edunet carried out through the eContentPlus Project, additional projects like Organic.Balkanet (www.organic-balkanet.eu) and CerOrganic (www.cerorganic.eu) contributed to the network with metadata, actual content and additional translations of the user interface of the portal and the metadata AP. In addition, the various content sources brought a number of educational resources in various languages with their metadata manually translated in several languages; however, this task proved to be too time-consuming and error-prone. In short a perfectly tailored portal with one exception: The multilingual support capabilities of Organic.Edunet were proving too cumbersome to users. This ran the risk of excluding potential participants and ultimately damaging the forward momentum of this hugely successful portal. It is here where Organic.Lingua came in.

2.2 After Lingua

Through the work of the Organic.Lingua project, Organic.Edunet has been transformed by the introduction of a range of additional capabilities into a truly multilingual experience. To date the portal interface is available in seventeen (11) languages. The Organic.Edunet Web portal currently provides access to resources in twelve (12) languages which are described with metadata in up to six (6) languages. New automatic translation components have been introduced in the revised version of the Organic.Edunet Web portal, in order to further facilitate the access to its content from various users.

In addition to the automatic metadata translation capabilities brought to the portal by the Organic.Lingua project, the team also focused on expanding the role of user-generated contributions to broaden both the scope and richness of the resource collection. In particular, Organic.Lingua focussed on enabling two types of contribution:

1. **Paradata:** Paradata allows users of the portal to get feedback from the community about the usefulness, accuracy or relevance of a given resource through the provision of ratings and qualitative text-based feedback.
2. **Metadata:** The users can contribute their own metadata in several ways
 - a. suggesting new resources to be available through the portal, by creating basic metadata records for these resources;
 - b. suggesting revisions in the existing metadata records by using a simple form for including these revisions and
 - c. suggesting translations for existing metadata records by providing their translations through a user-friendly form.

Organic.Lingua has achieved these advanced functionalities through the introduction of a new User Generated Content (UGC) widget, which is used for creating simple metadata records, compliant with the Organic.Edunet IEEE LOM AP. The widget provides the user with a simple interface featuring a small subset of the metadata elements available through the Organic.Edunet AP. These records are stored in a special repository and they go through a specific metadata management workflow which includes their evaluation and validation by a team of experts before they are published through the portal.

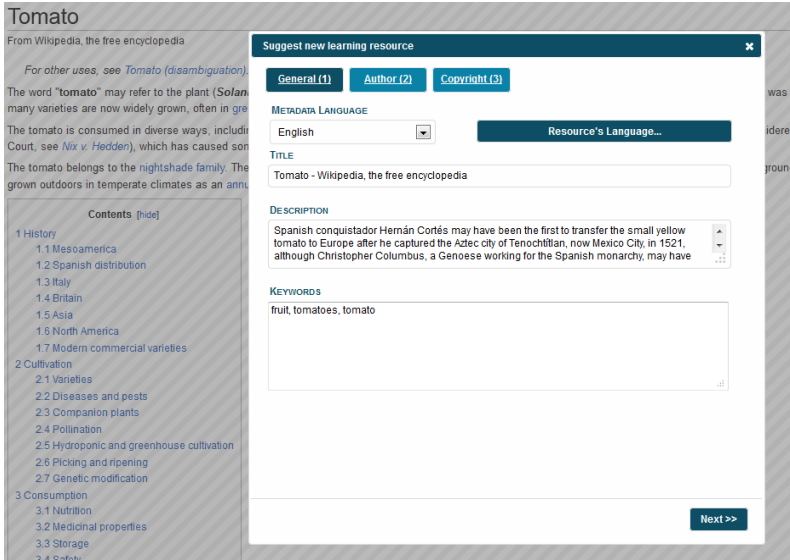


Figure 2.1. The UGC widget in the Organic.Edunet Web portal

In addition to the improvements developed through the User-Generated Widget, users are now able to customise their Organic.Edunet browsing experience through the form of personalised landing pages on the network. The team found this functionality was not only crucial in speaking to the different priorities of educational communities at all levels but was also incredibly beneficial as the development allowed Organic.Edunet to be presented not only as a multi-lingual portal but to provide the native language of user communities at the very first point of entry.

In addition, the user can refine the results that he is receiving after a query by using filters based on the metadata elements used for the description of the resources, including the language of the resource, the resource type (e.g. presentation, lesson plan, educational game etc.), the resource format (e.g. video, image, document) and the intended target audience.

Together the developments made during Organic.Lingua have made a useful niche content collection for those interested in the narrow educational sector of Organic Agriculture into a European entry-point of choice for educational materials in green education. The technical advances in translation capability developed during the project, together with a concerted

effort to broaden and advance the available content selection, have paid dividends for Organic.Edunet.

2.3 Revamping the Organic.Edunet Web portal

The Organic.Edunet Web portal acts as the user interface of all outcomes of the Organic.Lingua project. While there has been a lot of effort from both content and technical partners of the project, a user of the portal can only see the user interface of the portal without knowing the effort behind the development, adaptation and integration of a high number of different components. The Organic.Lingua project provided the opportunity for a complete revision of the Organic.Edunet Web portal and enhancements in several aspects, such as the user interface, the functionalities related to multilinguality as well as the enhanced contribution options for the registered users of the Organic.Edunet Web portal. The main enhancements are summarized in the following sections; however, more information about each one as well as additional ones are described in the following chapters of this document.

2.3.1 Multilingual content discovery

One of the biggest advantages of the Organic.Edunet infrastructure is its multilinguality. There are a number of features which support the multilinguality of Organic.Edunet, including the following:

1. **Multilinguality of the resources:** The portal provides access to learning resources in fourteen (14) languages, including several European ones, Hindi and Arabic. The integration of new collections is expected to enhance the multilinguality of the resources by introducing content in new languages.
2. **Multilinguality of the metadata:** The portal provides access to metadata in up to eighteen (18) languages, while additional translations may be automatically provided by the automatic metadata translations tools. A significantly high number of these metadata have been manually provided by humans and validated by language experts.
3. **Multilingual content retrieval:** With the integration of the Cross-Language Information Retrieval (CLIR) technology in the portal, the users may use search terms in their language and receive content in additional languages which they may also be familiar with. This functionality significantly enhances the user experience from the portal by providing a higher number of related resources to a specific query.

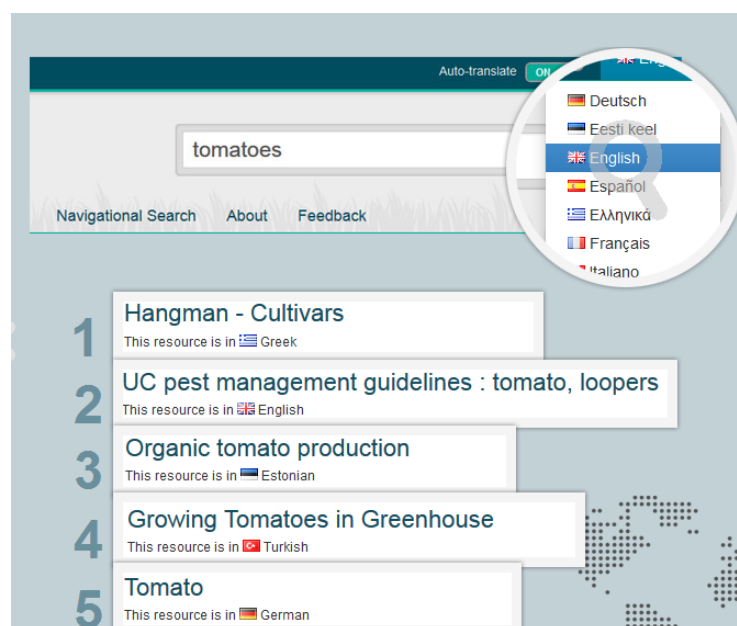


Figure 2.2: Overview of the multilinguality of the Organic.Edunet resources

4. **Multilingual user interface:** Another feature which improves the user experience from the portal is the fact that users can make use of the portal's interface in twelve (12) different languages. In addition, the metadata can be also displayed in the same language as the portal's interface, if the user selects so.

2.3.2 User evaluation of automatic translations and improvement

The role of the registered users in the Organic.Edunet has been enhanced; now, the users have the option not only to evaluate the automatic metadata translations provided by the machine translation components, but also to be more actively involved by revising these translations and suggest their own version of the translation.

More specifically, the users can evaluate an automatically-provided metadata translation based on a five-star scale. Apart from that, they can improve this translation by clicking on the corresponding button next to the language of the metadata and provide a revised translation using a simple user form. After the revised translation is evaluated by language experts of Organic.Edunet and is found correct, it is validated and published through the Organic.Edunet Web portal.

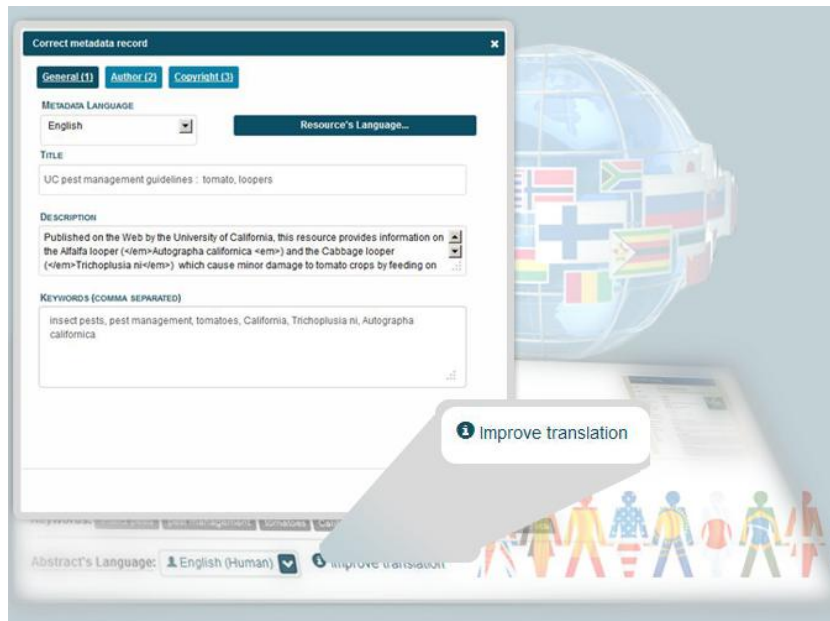


Figure 2.3: User-provided metadata translation

2.3.3 User generated content

The contribution of the portal users is not limited to the evaluation, rating and revision of translations. In the revised version of the Organic.Edunet Web portal, the registered users are also allowed to actively contribute to the content available through the portal by suggesting new resources.



Figure 2.4: The basic steps of the User Generated Content workflow

In practice, this is achieved through a bookmarklet that the users will have to add to their browser, just like a simple bookmark. Then, as soon as they find an interesting resource online, also appropriate for Organic.Edunet, the users only have to click on the specific bookmarklet. Through the use of a simple form, users are asked to provide a basic set of metadata and save the form. The new resource is evaluated by domain and metadata experts of Organic.Edunet and if it is found appropriate for publication, it is validated and published through the Organic.Edunet Web portal.

2.3.4 Domain-powered content discovery

What makes the Organic.Edunet Web portal special is its focus on organic agriculture, agroecology and other related green topics. In order to achieve this, the Organic.Edunet network consists of content providers with quality educational resources on the aforementioned topics. Before each collection is connected to Organic.Edunet, its content is evaluated by domain experts of Organic.Edunet using a pre-defined set of quality criteria. Additional evaluation in terms of the quality of the metadata takes place by metadata experts and if everything is ok, then the new collection is published through Organic.Edunet.

Periodic evaluation of metadata and content takes place after the collections are published through Organic.Edunet and in case issues related to the quality of either metadata or the content itself are identified, the corresponding content provider is contacted in order to take action.

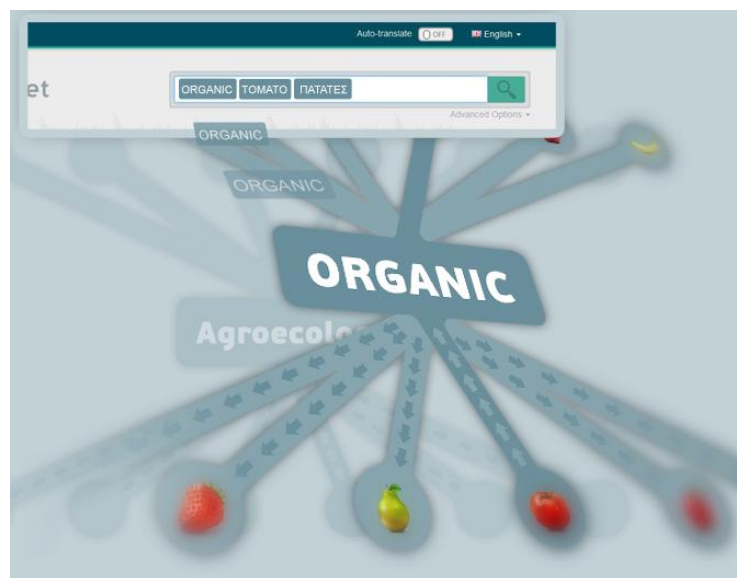


Figure 2.5: Domain-specific content retrieval in Organic.Edunet

However, the outcomes of the Organic.Lingua project are not limited to the aforementioned ones. The following sections provide information about additional Organic.Lingua outcomes and include some codified key lessons from the process so that others might benefit from our advances in the creation of their own multi-lingual content collections. They refer to individual components developed by the Organic.Lingua partners or adapter in order to meet the specific requirements of the Organic.Edunet Web portal users.



3 Enabling Multi-Lingual Content Search – Lessons for the Future

3.1 Overview of Organic.Lingua CLIR solution

The Organic.Edunet Web portal aggregates various collections of documents with their metadata available in different languages and offers its services to a multi-lingual audience, allowing users to search contents in their native language and retrieve results in all the available languages. Therefore it can constitute a useful reference example for anyone wishing to introduce language technologies into their portal.

In order to enable multi-lingual content search, different strategies can be adopted (*e.g. documents' translation, query translation, controlled vocabularies*) and in the Organic.Lingua project we opted for a Cross Language Information Retrieval (**CLIR**) system that performs user queries translation with a combined usage of bilingual dictionaries and domain ontologies; the systems then exploits the translated and semantically enriched queries in order to retrieve documents relevant to the queries.

The CLIR system implemented within the Organic.Lingua project is based on the following components:

- a Language Identifier tool, in order to identify the language of a textual fragment (the search query, a metadata field) when it is not known;
- a software tool for Lemmatization that is capable of selecting the base form of a word starting from its morphological variants and needed to retrieve the base form of the words (i.e. its lemma) used in dictionaries;
- a set of Bilingual Dictionaries for the supported languages;
- a set of domain knowledge bases, such as multilingual thesauri expressed in RDF, a language compatible with Linked Open Data technologies, which in our view enable interoperability of the informative assets between the partners and other stakeholders as well;
- a software tool capable of locating domain terminology elements from the supported ontologies within textual fragments; this component is used both at indexing time and at query time in order to annotate documents/queries;
- a Search Engine for actually searching the documents using the multi-lingual and semantically enriched query (in Organic.Lingua the search engine adopted was the well known open source SOLR¹).

Query submitted by users are analyzed in order to find labels pertaining to ontological concepts within the supported domain ontologies; the language independent URIs of the

¹<https://lucene.apache.org/solr>

found concepts constitute the multi-lingual semantic component of the query. In order to support this semantic component and allow the match between queries and documents, the index used at search time must contain references from documents to the supported domain ontologies either by means of human annotations or by means of the same automated process performed for query analysis. The quality of the semantic annotations within the search obviously impacts the effectiveness of this component and human annotations from domain experts should be preferred (when possible) to automatic strategies based on textual matching. In this way the supported ontologies actually act as controlled domain specific dictionaries.

The same query is as well translated by means of Lemmatization tools and Bilingual Dictionaries in order to support the topics (or the languages) not explicitly covered in the domain ontologies. The translations of the submitted query in the supported target languages constitute the multi-lingual terminological part of the query that is used as well to search documents within the index. The quality of the language resources used obviously impacts the effectiveness of this component, however, since these language resources tend to be expensive to be built and maintained, their quality should also relate to the degree of support desired for a given language.

The documents retrieved by the CLIR and presented to the user in the portal can then be translated to a target language by means of a Statistical Machine Translation (**SMT**) system (see section “Creating a Domain-Specific Translation Tool”).

3.2 Lessons Learned – Content Search

3.2.1 Plan the proper degree of support for the selected languages.

The first lesson learned in the implementation of such a system is the importance of selecting the languages that should be supported and with which degree of support. Even if the portal wishing to enable multi-lingual content search already exists, with an active base of users, it is indeed important to perform a statistical analysis of the portal search logs in order to understand which are the languages more frequently adopted by users in their interaction with the portal, and plan the right degree of support for those languages. For instance, a portal might be satisfied in supporting a certain language only with the domain specific ontological resources (or with a limited bilingual dictionary resources, e.g. having translations via a bridge language for some language pairs) if few users adopt it in their searches while requiring full support for another one.

3.2.2 Use High Level Interfaces to Wrap Language Resources.

Another important lesson learned in the course of the Organic.Lingua project is the importance of encapsulating the access to the language resources into high level interfaces in order to ease the maintenance of the different components and allow for a seamless integration of new resources. This is particularly important in order to ensure the modularity

of the system and allow for the introduction of a new language or increasing the support of an already present one.

Access to language resources should be as well injected into the system by means of configuration properties in order to allow simple modification to its functionalities (e.g. changing the path of a static resource or the URL of a web service).

3.2.3 Reuse Public Knowledge

Another lesson learned is on reusing public knowledge whenever it is possible. Publicly available knowledge on a wide variety of domains exists on the web and by supporting the proper standard format it is possible to include and reuse it in any portal. Within the Organic.Lingua project we adopted the RDF SKOS representation format, and included the well known AGROVOC² Thesaurus from FAO, which contains about 32,000 concepts in up to 22 languages. The main advantage of integrating public knowledge sources consists in having good quality resources as domain dictionaries for free and without maintenance costs; other advantages arise when they are coupled with knowledge resources specifically developed for the portal.

3.2.4 Link the Portal Knowledge to the Public Knowledge

Available public knowledge might cover only partially (or not with the desired degree of precision) the topics exposed in the portal contents. Whenever a knowledge resource specifically developed for the portal contents exists, it is a good practice to map (either manually or via automated procedures) its concepts to the ones of a well-known public knowledge source. In this way the concepts present in the portal specific resource can be enriched with new synonyms or new languages. Organic.Lingua ontology covers 11 languages and, by mapping it to AGROVOC (that covers 22 languages), we obtained an important increase of the languages supported by the domain terminologies.

Besides the support of new languages for free, the richer lexicon obtained with the mapping between Organic.Lingua ontology and AGROVOC enhanced as well the effectiveness of the information retrieval system for the previously supported languages as confirmed by internal evaluation tests performed within the project.

3.2.5 Involve Users

Another lesson learned consists in involving users in the process of maintaining/refining the language resources (e.g. by providing a better translation for a given term) in order to allow the system to take advantage of user contributions and enrich specific language resources with the growth of given language speaking audience of users.

²<http://aims.fao.org/standards/agrovoc/about>



3.2.6 Caching Mechanism

The semantic enrichment of the query and its multilingual translation are computationally expensive operations, therefore the system might benefit from caching mechanism in order to avoid duplicating the same request, if the language and knowledge resources used in the process (and thus the final result) would be the same.

4 Enabling Machine Translation- Lessons Learned

4.1 Overview of *Organic.Lingua MT solution*

The Organic.Edunet Web portal offers its services to a multi-lingual audience, allowing its users to view the contents of the portal in a number of languages supported by Machine Translation tools that have been adapted to the Organic.Lingua domain.

In order to enable translation of the portal's contents, several strategies could be pursued, such as rule-based machine translation or statistical machine translation (SMT). SMT systems have several advantage over rule-based systems in that they are less costly to develop, can capitalize on existing open resources such as generic bilingual corpora, and can be adapted to new domains by exploiting in-domain corpora, that is, corpora that address topics close to the actual texts to be translated.

The SMT systems that have been developed by Xerox specifically for the needs of Organic.Lingua consist of the following main components:

- A bilingual phrase table that provides translational equivalences between sequences of words in the source and target languages.
- A language model that provides an assessment of the quality of a sentence in the target language.
- A decoder which takes as input a sentence in the source language, uses phrases from the table to assemble a candidate target sentence, and searches for the best target sentence, based on different characteristics of the phrases, on the assessment provided by the language model, and on weights for combining the characteristics and the assessment.

In order to obtain the phrase table, the language model, and the weights, a training process is necessary. This process requires the availability of bilingual corpora, consisting of sentences aligned with their (human) translation, as well as monolingual corpora used for producing the target language models. While it would be optimal to train on large "in-domain" corpora (corpora similar to the texts that should be translated), in practice such corpora are small and difficult to find. What we do instead is to train a first model on large (easy to find) generic corpora, and to use the smaller in-domain corpora as additional data for "adapting" the first model to the domain to be handled.

A detailed description of the SMT components and of the adaptation methods can be found in the Deliverables "D4.1.1" and "Report on Domain Adaptation of linguistic services".

4.2 Lessons Learned – Machine Translation

4.2.1 Plan the proper degree of support for the selected languages

In common with other language components such as CLIR, it is important for the machine translation components to identify at the outset of the project which languages should be supported and with what degree of effort. This should be done on the basis of usage statistics. For example, if speakers of a language L2 often require access to contents in a language L1, then it is worth spending a lot of effort on adapting a generic translation model L1 -> L2 to the specific domain of the portal, whereas if the language pair is more rarely needed, then it may be sufficient to use a generic translation model.

4.2.2 Provide careful estimates of the effort required for in-domain adaptation

Domain adaptation of a generic translation model for the specific needs of a portal is a process whose difficulty should not be underestimated, and for which a significant effort should be planned. This concerns several aspects. First, one needs to identify as precisely as possible the nature of the texts that will require translation, which may actually be a mix of different domains (e.g. Agriculture and Education). Second, one needs to identify available bilingual and monolingual corpus as well as terminology resources associated with these domains. Third, one needs to extract these resources, clean them, and provide them in the common format required by the training modules of the underlying translation system; one needs to balance the effort required to incorporate a new resource with the expected gain on the translation performance. Finally, one needs to experiment with different combination of adaptation techniques in order to get optimal performance, an endeavor which is currently more an art than a science.

4.2.3 Identify the test sets carefully at the beginning of the project

In order to provide measures of the evolution of the performance of a system along the course of the whole project, it is necessary to choose, early on, test sets of human translations on which the quality of the systems will be evaluated; while these automatic evaluations do not replace the human evaluations that provide a more important assessment of the overall usability of a translation system, the automatic measures have the advantage of being cheap to produce and of providing objective measures of the incremental improvements gained in the course of the project. It is however important to choose these test sets carefully to be close to the nature of the texts for which a translation will actually be needed, otherwise the automatic measures may not be strongly correlated with the actual perceptions of the users.

4.2.4 Anticipate risks and have fall-back plans

Because of the costs associated to the development of domain-adapted translation components, it may be necessary to exploit externally provided generic translation services. However, the availability and cost of these systems may evolve in unpredictable way; for instance, in the course of the Organic.Lingua project, some well-known such services have moved from free-to-use API to a pay-to-use model. It is then advisable, before the beginning



of the project, to take measures for such eventualities. One way to do so is to plan to use, as much as possible, open-source translation systems, and to plan the (relatively small, but not negligible) costs of deploying these systems for use by the portal.

4.2.5 Use High Level Interfaces to Wrap Language Resources.

Similarly to the case of CLIR and language services in general, another lesson learnt is the importance of encapsulating the access to the different translation services into high level interfaces in order to ease the maintenance of the different services and allow for a seamless integration of new ones.

4.2.6 Caching Mechanism

Translating texts can be a costly operation, and the available servers may not always be powerful enough to respond quickly enough to ensure an optimal user experience. However, one observes that it is often the case that the same text needs to be repeatedly accessed, and in such cases a caching mechanism for already performed translations is an effective way to speed up the user experience. In the limit, only texts that are accessed for the first time need to actually call the available translation services.



5 Enriching your Metadata – Lessons for the Future

5.1 Overview of the Organic.Lingua Metadata & UGC System

The Organic.Edunet network is based on the sharing, harvesting and publication of educational metadata records describing the corresponding educational resources. In some cases, the collections of the Organic.Edunet content providers consist of poor and/or incomplete metadata, not translated in additional languages. This situation may raise barriers for users wishing to access the specific resources, as a complete metadata description significantly facilitates access to the actual resource and its retrieval. The use of enriched metadata records provide additional ways of retrieving the content facilitating search mechanisms based on specific metadata elements (such as language of the resource, resource format and type, educational level, intended audience etc.). Additional classification of the resources based on the Organic.Edunet ontology further facilitates the indexing and retrieval of the resources.

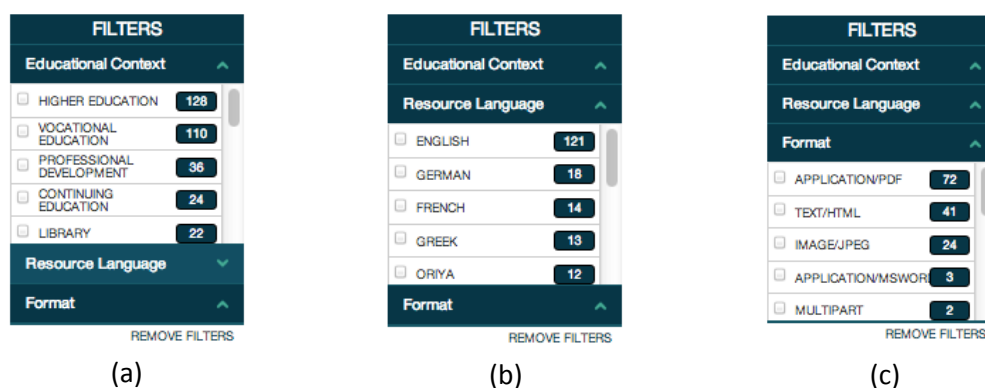


Figure 5.1: Some of the available search filters based on educational context (a), resource language (b) and resource format (c)

Content providers can join the Organic.Edunet network and publish their metadata through the Organic.Edunet Web portal through various ways, including the following:

- **Harvesting:** Metadata repositories supporting the OAI-PMH standard can be directly harvested by the Organic.Edunet Metadata Aggregator;
- **Ingestion:** Educational collections which do not support the OAI-PMH standard but still have metadata which can be exposed as files (e.g. XML/CSV dump), can share their metadata with Organic.Edunet;
- **Manual indexing:** In the case of content providers who have educational resources in the scope of Organic.Edunet but do not have access to a repository tool, they can use the Agricultural Learning Repository Tool (AgLR) in order to create a metadata collection and have it published through the Organic.Edunet Web portal.

In all cases, it should be noted that the new collections should meet a set of basic criteria regarding the quality of the provided resources and their metadata, their relation to the thematics covered by Organic.Edunet etc.

In addition, users can contribute to the content available through the Organic.Edunet Web portal through various ways, as shown in Chapter 2. Additional means for user contributions include:

- **Rating resources:** The user can rate a resource using a 5-star scale, regarding the quality of the metadata, the relation of the resource with the topics covered by Organic.Edunet as well as its educational usefulness.
- **Reviewing resources:** The users can write a short review for each resource, which will be available to all other users of the portal. The review could highlight the strengths and weaknesses of a resource, as well add related information.
- **Tagging resources:** The user can tag a resource with related keywords, in order to make it even more accessible to other users, while at the same time this makes it easier for him to locate the tagged resources.

In this way, the resources provided to the Organic.Edunet network by a content provider receive an added value provided by the users of the portal.

5.2 Lessons Learned – Metadata Enrichment

In the context of the Organic.Lingua project there was a number of new collections introduced to the Organic.Edunet network with their metadata integrated through various ways (e.g. harvesting, ingestion and manual annotation). Each case exhibited different attributes and issues, which were addressed on an individual basis. This experience can be summarized as follows:

5.2.1 Even automatic metadata harvesting/ingestion can be problematic

In the case of collections which were directly harvested through the Organic.Edunet aggregator, in several cases there were issues related to the correct implementation of the OAI-PMH protocol in the repositories. In most cases this issue was addressed after communication with the content providers and technical support by the Organic.Edunet aggregation team. In the case of ingestion, in various cases the metadata records were not properly formatted/structured, while information considered mandatory by Organic.Edunet was not available in the records. On top of that, transformation of the metadata schemas used by the content providers to the Organic.Edunet IEEE LOM AP was sometimes tricky, as the semantics of the source schema were not well-defined. The completeness of the metadata records was also an issue, especially in cases where the missing information was defined as mandatory by the Organic.Edunet IEEE LOM AP.

5.2.2 The use of the appropriate metadata authoring tool can help metadata authoring

However, the most important lesson learned came from the manual annotation of educational resources with metadata using the AgLR tool, as well as the enrichment of existing metadata with additional information which was available. A process which used to be time-consuming for the metadata author became less cumbersome thanks to the automatic metadata extraction functionality of the AgLR tool. The time needed for the translation of metadata was also reduced, thanks to the automatic metadata translation functionalities integrated in the AgLR tool. Last but not least, the time required for the creation of metadata for resources bearing similar characteristics (e.g. same author/publisher, same keywords, educational characteristics, classification terms and IPR descriptions) with the use of the Template functionality of the AgLR tool. The AgLR tool is also connected with the Ontology Service API, so it has access to the latest version of the Organic.Edunet ontology at all times through an automated way, compared to the manual update of the ontology in the tool, which was the case in the context of the Organic.Lingua project.

5.2.3 Multilinguality matters

There is already a wealth of information and educational resources related to organic agriculture, agroecology and similar topics available through various Web sources; however, this information is not always available due to linguistic barriers. For example, metadata in a language not spoken by a user cannot be of any help for him, even if they describe an image – this also means that it will not be possible for a user to retrieve this resource by using English terms for retrieving resources. However, in the context of the Organic.Lingua project, a relatively high number of resources were annotated with multilingual metadata, a fact which increased the availability and access of these resources to a wider audience. This is facilitated by the use of the AgLR tool, as the change of the language of the tool is not limited to providing a fully translated user interface but also changes the language of the metadata schema and vocabularies in the same language. Being multilingual is a clear advantage.

5.2.4 High quality metadata need time and guidance

As mentioned earlier, not all content sources of Organic.Edunet provide high quality metadata. In some cases even information which is easily accessible is not provided in the metadata records and in extreme cases, the metadata records only included title and some keywords (skipping the description and information on the rights applying to a resource) and lacked any kind of classification terms. In other cases, there were collections where the formatting of the metadata information was really poor (e.g. symbols appearing in the title, metadata was created in a language with specific characters which were not correctly transformed/displayed, such as accented characters etc.). In case of small collections, collection managers were guided into revising and improving the quality of their metadata before they were ingested by Organic.Edunet. In the case of incomplete metadata, missing information was semi-automatically inserted in the records as long as the same value

applied to a whole set of metadata (e.g. name of publisher, classification terms, vocabulary values etc.). In all cases, the identification of the issues (through the metadata quality evaluation and technical testing of the OAI-PMH targets) and the selection of the most appropriate solution needed planning and careful design in order to ensure that the issue would be fully addressed. In addition, the documentation and guidelines available for the use of the Organic.Edunet IEEE LOM AP were helpful, along with the availability of the Organic.Edunet data team, which provided help and guidelines to the content providers when needed. Quality is a priority over quantity when it comes to metadata in Organic.Edunet.

5.2.5 Use the power of the community

Organic.Edunet focuses on the quality of the provided resources over their quantity. Content sources are carefully selected and go through an evaluation process, which includes evaluation of the quality of their metadata. However, it is not possible for the Organic.Edunet network to keep updating the resources available through the portal, ensure that all content is appropriate and identify issues in the metadata. This aspect can be addressed with the engagement of registered users of the portal, which are provided with additional functionalities allowing them to be actively involved in the kind and quality of the content available through the portal. In this direction, registered users of the Organic.Edunet Web portal can suggest new content to be published through the portal, can propose new metadata translations and report inappropriate resources available through the portal.

5.3 Connection with GLN

The Green Learning Network (GLN) is a large network of content providers with educational content for agriculture in general. The topics covered by GLN are much wider compared to Organic.Edunet as well as the number of the content providers. In this direction, GLN acts as a pool of potential new collections/content providers for Organic.Edunet, since all content related to organic agriculture, agroecology and the rest of the topics covered by Organic.Edunet can be published through the Organic.Edunet Web portal.

Organic.Edunet is the organic agriculture/agroecology branch of GLN and benefits from it, as GLN acts at a larger scale, consists of a higher number of content providers and has access to opportunities for networking even more content providers, collections and repositories. In this way, GLN can also contribute to the sustainability of Organic.Edunet, as the effort needed for maintaining and expanding the Organic.Edunet network can be at least partially covered by the corresponding GLN activities.

6 Engaging your User Base – Lessons for the Future

The preceding sections of this white paper have all dealt exclusively with the technical advances made during the Organic.Lingua project and their positive impact on the user experience. However, it must be noted that technical advances, no matter their merits, have no value if the intended beneficiaries of your portal are unable or unwilling to join in and use your solution. To counteract this danger, the Organic.Lingua team engaged in a full programme of public promotion and trials to ensure that new features were quickly associated in the minds of users with Organic.Edunet and that good practices of platform use were promoted by ‘champions’ such as teachers or facilitators across Europe.

In addition to the performance of closed user-group testing in each pilot country of the project, Organic.Lingua found it beneficial to conduct Open Public Trials of the solution online using the existing Organic.Edunet user base as a rich evaluation source already familiar with the resources available and in a unique position to interpret the additional benefits and functionalities provided by the Lingua solution.

Once of the points that was especially important for the communication and promotion team of Organic.Lingua was not to simply focus on the translation improvements offered by the project but to also illustrate the improved user experience of the Organic.Edunet portal through the multilingual functionality. A concerted marketing push toward this message of improved functionality paid dividends in terms both of user satisfaction and user numbers.

6.1 Lessons Learned – User Base

6.1.1 Get Influencers Onside Early

One of the key priorities for the Organic.Lingua team was to quickly engage influencers in the relevant communities to champion the new multilingual functionality. Hearing an endorsement from a trusted figure in the domain is one of the most powerful ways to engage a user and this lesson is particularly apposite when considering translation issues. Organic.Lingua focussed on engaging with educational thought leaders early to get them championing the solution to their own communities and building the user base through an organic network.

6.1.2 Identify the target audience

It is important to get to know the target audience of a tool, like a learning portal in the case of Organic.Edunet. This will allow the development of custom services which will make the portal more attractive to its users and will increase the effectiveness of any dissemination and exploitation activities as they will be targeted to specific audience (mailing lists, community fora, individual contacts, organizations etc.) instead of promoting through generic channels. In the context of the Organic.Lingua project the target audience was clearly defined and this allowed the development of domain-specific solutions, such as the



domain-specific metadata translation tools, the Agricultural Learning Repository Tool (AgLR), the adaptation of an ontology management tool (MoKi) etc. In addition, when the feedback from the (potential) users of the portal was needed, the people who might be interested in testing the Organic.Lingua outcomes and providing feedback were contacted through valid mailing lists, related organizations and initiatives.

6.1.3 Keep the stakeholders engaged

Another important aspect is to keep the communication with the target audience alive and preferably on a periodic basis. Identifying opportunities for communication is important and in the case of Organic.Edunet could include updates related to the Organic.Edunet Web portal, circulation of the Organic.Edunet newsletter, information (as well as invitations) to Organic.Lingua related events, invitation to test new functionalities of the Organic.Lingua outcomes and share provide their feedback etc. The stakeholders need to feel that they are valued and that their feedback is indeed taken into consideration, so sharing news about the results of a survey or the updates on the portal based on the feedback received is always highly appreciated.

6.1.4 Tie-In Events Provide Real Hands-On Feedback

Many projects make the mistake of leaving user evaluation until the final phase of the project in a formal, structured set of workshops. This approach runs the risk of identifying usability challenges too late. Organic.Lingua found that it was helpful to have a series of informal feedback and testing sessions at creativity events such as Green Ideas³ to provide early opportunities to get hands-on feedback from users and iteratively refine and improve the product along with constant feedback.

³ <http://greenideasproject.org/>

7 How to Use the Organic.Edunet Solution

7.1 *Initial steps*

The Organic.Edunet Network is always in search of new collections to be interconnected and published through the Organic.Edunet Web portal, so it is highly likely that a content provider will receive an email from the Organic.Edunet team, asking him to investigate the possibility of collaboration in this direction. However, all content providers with educational resources covering green topics are more than welcome to contact the Organic.Edunet team in case they are interested in joining Organic.Edunet and publishing their metadata through the Organic.Edunet Web portal.

- The first step is the initial communication and the acknowledgement of the new collection to be connected to the Organic.Edunet Network.
- Information about the collection and responsible organization will be requested using a Registration Form for New Content Providers.
- The Organic.Edunet team will ensure that the new collection is complying in terms of appropriateness and quality of resources and metadata, ensuring at the same time that the Organic.Edunet Pre-Check Core Criteria are met and accepted by the content provider.
- A simple and descriptive Memorandum of Understanding (MoU) in the form of a Data Exchange Agreement (DAE) is signed by both parties, clarifying the most important aspects of the collaboration.
- More detailed information about metadata schema used by the content provider may be requested by the Organic.Edunet team using the Collection/Repository Registration Form. This information will be used for the mapping of the metadata records to the Organic.Edunet IEEE LOM AP.
- The optimal way of collecting the metadata records is selected (e.g. harvesting, ingestion), according to the technical capabilities of the content provider's system.
- After a testing & calibration phase, the new collection will be ready to be published through the Organic.Edunet Web portal.

7.2 *Options for publishing content through the Organic.Edunet Web portal*

1. **Harvesting option:** This refers to content providers that have a digital collection described with metadata and organized in a repository tool. If the tool supports exposure through an OAI-PMH target, then the metadata can be automatically harvested, mapped to the Organic.Edunet metadata AP and then exposed to the



Organic.Edunet Web portal. If a target is not set up, then the metadata can be exposed as XML or CSV files, mapped to the Organic.Edunet AP and then exposed through the portal.

2. **Ingestion option:** This option refers to the creation of metadata through various automatic ways, such as tools that connect to collections of metadata using the corresponding API and create Organic.Edunet AP-compliant metadata records. This is the case of sites like YouTube, Flickr and Slideshare, which are excellent sources of multimedia content within the scope of Organic.Edunet. In addition, crawlers collecting information for the creation of metadata records based on information already available in the website of a content provider also fall under this category.
3. **Manual Indexing option:** If the collection to be connected to the Organic.Edunet network is not organized in any kind of digital tool, then Organic.Edunet can propose a number of tools that can be used for this purpose (such as the AgLR tool). The content provider can then manually create a set of metadata records for his collection and automatically make it available through the Organic.Edunet Web portal.

8 Conclusion

The Organic.Lingua white paper set out with the simple objective of presenting some of the transferable lessons learned during the Organic.Lingua project that can be carried forward to other multi-lingual projects. To summarise, Organic.Lingua has learned the following key lessons during the project:

Enabling Multi-Lingual Content Search

- **Plan the proper level of support for your languages** – Always understand how much users rely on a given language. If you have 2000 searches a month in French and 20 in Russian, it makes sense to dedicate a greater amount of support to French.
- **Use High-Level Interfaces to Wrap Language Resources** – Always integrate resources in such a way that you can easily add new languages to the system.
- **Reuse Public Knowledge** – Don't reinvent the wheel where a public resource exists that will work for your project.
- **Link the Portal Knowledge to the Public Knowledge** – Where domain-specific ontologies can be found in the public sphere, integrate them with your own.
- **Involve Users** – Your users are closest to the translated content on a daily basis, enlist their help in managing the system
- **Caching Mechanism** – To speed up responses and avoid duplication, cache popular translation requests.

Enabling Machine Translation

- **Provide Careful Estimates of Effort for In-Domain Adaptation** – Understand and honestly assess the cost of adapting translation software to your specialist domain. The more special terminology, the higher the likely cost.
- **Identify Test-Sets at the Beginning of the Project** – Carefully select which language pairs will be used to assess the domain and start with a simple set of languages that can be supported most easily.
- **Anticipate Risks and have Fall-back Plans** – Domain translation is rarely perfect and problems will arise. What is essential is that the team envisage these possibilities and react with agility to challenges.

Enriching your Metadata

- **Even Automatic Harvesting can be Problematic** – Metadata harvesting requires careful supervision to ensure that the results prove a useful resource for users.
- **Authoring Tools can Assist the Production of High-Quality Metadata** – Wherever possible create or implement a tool that makes the lives of content-owners easier to improve the quality of your metadata



- **Multilinguality Matters** – Having good-quality multilingual metadata can create a dramatic improvement in the discovery and use of your content resources.
- **High-Quality Metadata needs Time and Guidance** – Create clear guidelines for your content owners and dedicate the time to making sure Metadata is useful and clear.
- **Use the Power of the Community** – Harness the collective power of your user base to find and report bad resources, improve translations and enrich metadata

Engaging the User Base

- **Get Influencers Onside Early** – Win over community champions to your project and they will promote the initiative through their own networks
- **Identify the Target Audience** – Understand very clearly who your audience is and always design with their needs in mind.
- **Keep Stakeholders Engaged** – Spread out content and announcements to keep your stakeholders, particularly users, engaged with the platform
- **Tie-In Events Provide Real Hands-On Feedback** – Use events and other opportunities to test the system early and often on real groups of people.

We are always happy to discuss our work with the Organic.Lingua project and welcome any opportunity to pass on our collected lessons to new projects. To connect with the team, simply visit <http://www.organic-lingua.eu/>

Annex: The Organic Eprints Discovery Space

As mentioned in the previous chapters of this document, the Organic.Lingua project outcomes provided the Organic.Edunet Web portal with a number of enhancements focusing on multilinguality but not limited to it.

Organic Eprints is one of the largest databases / repositories of resources on organic agriculture, focusing on scientific publications and research outcomes but also featuring other types of resources such as country-specific reports, teaching/training resources, description of related programmes and projects, journal articles etc. The Organic.Eprints portal provides basic search and browsing functionalities, based on the typical ones provided by the Eprints software. In addition, the user interface is only available in English and German; however, the analysis of the statistics of the portal revealed a high number of users and visits from France, Spain and Italy, among others.

In this direction, there was a collaboration between Organic.Edunet and Organic Eprints, through which the multilinguality features of the Organic.Lingua project could be applied to an external case and be used as a reference for further applications. A specific methodology was applied, which included several steps starting with the the collection of user requirements and ending with the development of a first version of the Organic Eprints discovery space. The analysis focused on the multilinguality needs of the Organic Eprints users and also took into consideration the resource types and thematics available in Organic.Eprints, the most populated categories etc. The methodology followed, as well as the Organic Eprints Discovery Space, are described details in the deliverable **D5.2.3 “Final multilingual deployment of the Organic.Edunet Web portal”**

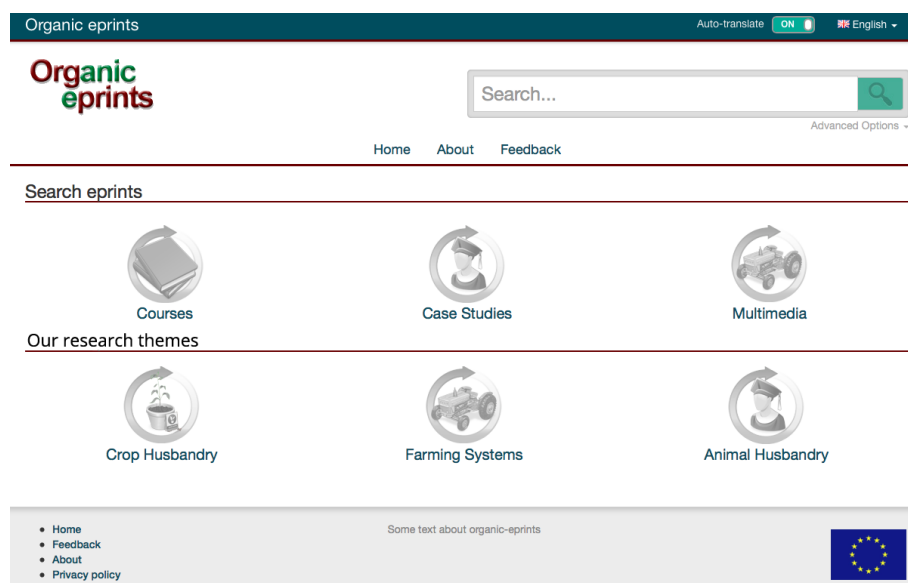
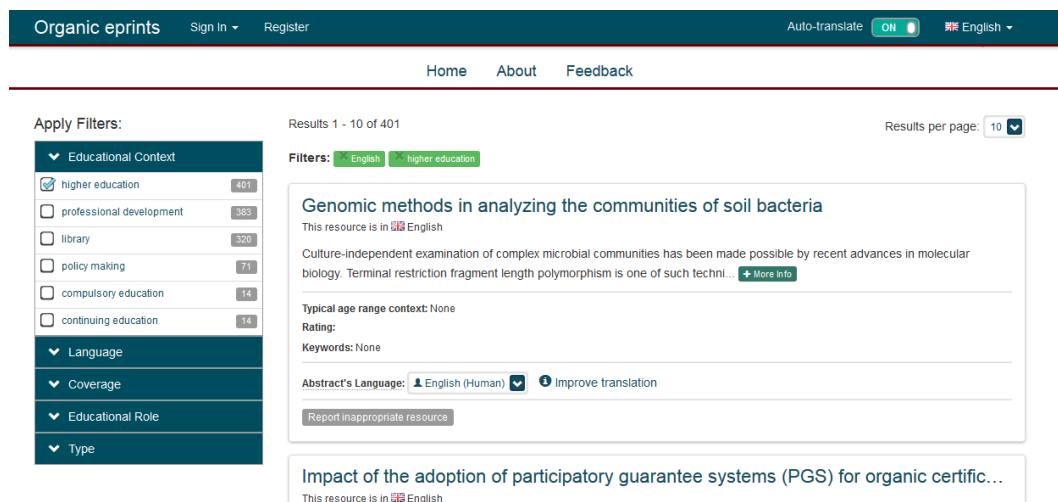


Figure: The Organic Eprints Discovery Service

The Organic Eprints Discovery Service is based on an adapted version of the re-engineered Organic.Edunet Web portal; in this direction, it offers the same set of multilinguality services to the users, such as:

- Automatic metadata translation and option for users to rate and edit/improve existing translations;
- Multilingual user interface;
- Cross-language information retrieval.



The screenshot displays the Organic Eprints Discovery Space interface. At the top, there is a navigation bar with 'Organic eprints', 'Sign In', 'Register', 'Auto-translate' (set to 'ON'), and a language dropdown (set to 'English'). Below this is a secondary navigation bar with 'Home', 'About', and 'Feedback' links.

The main content area shows search results. On the left, there is a 'Apply Filters' sidebar with categories: Educational Context (with sub-items like higher education, professional development, library, policy making, compulsory education, continuing education), Language, Coverage, Educational Role, and Type. The 'higher education' filter is selected, showing 401 results. The 'Filters' section at the top of the results area shows 'English' and 'higher education' as active filters.

The search results list two items:

- Genomic methods in analyzing the communities of soil bacteria**: This resource is in English. Description: Culture-independent examination of complex microbial communities has been made possible by recent advances in molecular biology. Terminal restriction fragment length polymorphism is one of such technol... [More info](#)
- Impact of the adoption of participatory guarantee systems (PGS) for organic certific...**: This resource is in English.

Each result entry includes fields for 'Typical age range context', 'Rating', and 'Keywords'. The first result also has an 'Abstract's Language' dropdown set to 'English (Human)' and an 'Improve translation' button. A 'Report inappropriate resource' button is located at the bottom of the first result's entry.

Figure: The metadata view in the Organic Eprints Discovery Space

In addition, the registered users of the Discovery Service can suggest content for Organic Eprints following the same workflow as in the case of Organic.Edunet. Last but not list, the Organic.Eprints Discovery Space offers filtering of resources based on pre-defined metadata values, such as the Language, Coverage and Type of resource.

The development of the Organic Eprints Discovery Space showed that the components, tools and workflows developed in the context of the Organic.Lingua project are reusable and can be applied in several other cases