Integrating true multilingual capabilities into an Institutional Repository

Building the World Health Organization's Institutional Repository for Information Sharing

Introduction

In a global context, how do we facilitate the dissemination and access if the material in a repository is primarily searchable and retrievable in only in one or two languages? It has been observed that there is much research and public health guidelines that goes unknown to large numbers of researchers, health workers and to the general public when they are only able to access in one language or another. How do we promote integration of various information sources in an international organization with 147 country offices, six regional offices and one headquarters, and with material being published in 6 official languages and 53 non-official languages? Research ethics should start considering, at design stage, the outreach of methods used and results obtained beyond the boundaries of the research language. Access to information in as many languages as possible should become a major component of any accessibility-related debate.

Background – Context

IRIS (Institutional Repository for Information Sharing) is a digital library of WHO's information products. IRIS was developed in response to WHO'S Executive Board and the World Health Assembly's multilingualism plan of action in 2007, which required a global institutional repository of WHO publications and documents.

The aim is to enhance the collection, storage and preservation, and to facilitate the dissemination of its technical information in the Organization's six official languages (Arabic, Chinese, English, French, Russian, Spanish), and in digital format.

Data will be stored in a decentralized model and common standards will permit full interoperability and searching across all data. Moreover, the repository will have a multilingual interface, its content will be openly and freely accessible and fully searchable in all six official languages, and the collections will include WHO information products (publications, governing bodies' documents, archives and scientific and technical reports.

An open source repository software, DSpace, was selected as the most suitable tool for IRIS for its flexibility in terms of multilingual interoperability, standards for preservation and dissemination, and integration with other digital libraries and bibliographic databases and the WHO website. DSpace is available in several languages and yet is not fully integrated to serve as a multilingual repository. Some expansion of its functionalities are required for IRIS.

Outline of DSpace customizations and multilingual capabilities incorporated into WHO IRIS.

There were several primary customizations to DSpace to achieve this:

- **MeSH lookup** for submission process and meta data update processes

- Customization of meta data views, including:
  - Handle to language mapping for record linking,
  - Displaying Mesh qualifier information,
  - ISO language code to **language name mapping**,

- Simple **meta data localization** of views, browsing and searching depending on the current locale. (disabled for Phase 1)

- **Language translation management** via a tool for importation of ~1500 labels for each of 6 languages, and exported as ascii encoded language properties files for use in DSpace application.

- Meta data extensions through the use of customized schema namespace. who.relation.languageVersion to be able to **link language versions** of the same document

Background and detail

Existing DSpace

DSpace embraces multilingual capabilities for its user interfaces - text strings throughout the UI are separated into message files that can be exchanged for the user's chosen language. Whilst the fixed labels and messages that form the interface are well catered for, there is less provision for the dynamic content of the repository - the metadata of items, the names and descriptions of the communities and collections.

Item metadata does have some provision for multilingual uses. Each piece of metadata is not only identified by the field that it is - it also has a language qualifier on it. By using an ISO language identifier, you can specify what language is used in that particular instance - e.g., in an item you could have two 'dc.title' instances, one marked as 'en' and one as 'fr'.

Although the language qualifier is used in some instances of the sorting code - i.e., in order to strip the definite/indefinite articles from the start of a title - it is otherwise ignored by DSpace. If you go to an item that has multiple language instances of a field, then you will just see all of them displayed on the item page.

Initial customizations

Due to change of project direction to emphasize translatable Authority Control and MeSH lookup, these first modifications were not completed and were shelved until a later Phase.

Meta data filtering based on user's language

Our **first modification** was to allow the item page to filter what metadata is displayed, based on the language qualifier and the user's chosen locale. So, if an item had 'dc.title' instances that were in English and French, and you had chosen the French interface, then you would only see the French instance of the title. It would also fall back to the default locale, in the case that the metadata wasn't available in your chosen locale - in the previous example, if you had chosen a Russian interface, then it would display the English title, as there wasn't a Russian title available.

Communities and Collections Meta data table extension to accommodate multiple languages

Our **second modification** was to take the communities and collections and allow them to work in the same way as the items, as described above. DSpace currently has all the metadata - the title, description, etc. - for communities and collections as columns in the database on the collection/community tables. This means that not only is it inflexible for adding more data to community and collection pages, you don't have multiple instances and language qualifiers to allow you to have multiple translations of the title, description, etc.

In order to allow for more flexible metadata descriptions, we extended the existing metadata table that is used for items, and allowed it to be related to communities or collections. This would allow us to re-use much of the metadata handling from items in communities and collections, and provide us a flexible means of describing communities and collections that would incorporate handling translations. By having metadata described in the same way as items, we would have been able to filter the multiple translations to only display the one relevant for the current language selection, just as we could for items.

This work was not completed or integrated though, as it was identified that the items would need to exist as separate records for each language, both for import purposes, and the possibility of interoperating with other local repositories that would only use one language.

Change of direction – Current Customizations

Linking different language versions

As multiple language versions would exist in the repository as multiple items, a way of linking the different language versions was required. Without an identifier that would allow us to automatically link to other version of the same document, creating links between the items would have to be a manual process of entering the ids of related items into the metadata.

Our **third modification** was to use the identifiers of related versions to find out what language those items are defined to be in, and use that information to present understandable links to the end user as to what other languages are available for that item.

Authority Control, MeSH, Automated BabelMeSH translation

One other requirement was also identified - that MeSH terms should be presented to users in their chosen language, so that they could navigate to items related to the subject that interest them, without specifically having to know the English terms. In order to do this - and to ensure the accuracy of the MeSH terms - we placed the MeSH metadata entries under 'authority control'. This authority control is a new feature of DSpace 1.6 that validates the entry against known terms, and we extended this to accommodate MeSH lookp and translation.

Authority control had some provision for keys (values entered into the metadata fields), and labels (values displayed to the user), and it already allowed for locale information to be used in the choice of labels. This was not consistently implemented throughout DSpace though.

Our **fourth modification** was, therefore, to make the use of keys and labels consistent. Labels would now always be used when viewing item records, or the browse list entries. By using the labels throughout the interface, we could pass the user's chosen language through and get the translation of the MeSH term into the current language (providing it was translated in BabelMeSH).

In doing this, we also modified the browse code so that it could sort the translations of the authority controlled fields (MeSH) correctly for the chosen language. The search code was also modified so that it would index the different translations of the MeSH, allowing it to be searched in a user's native language.

Additional aspects of multilingualism – translation and user interface

Management of 6 official languages

The interface of the IRIS required that the 1500 labels that make up an instance of DSpace be translated into the 6 official languages of the WHO. DSpace comes with English and even this needed modification to more accurately portray the repository as WHO IRIS rather than the default wording of DSpace. This was modified by WHO staff and imported into a tool for managing java language properties files. This tool is called Tongue Tied and was modified by BMC to be able to handle the 6 official languages, each of which has its own challenges, that of character sets, and right to left for the Arabic. There were some translations available for DSpace, but all were incomplete and some hardly at all. The Tongue Tied tool allowed for the export of the languages in columns side by side in Excel and this excel sheet was distributed to WHO translation teams and then reimported back into the tool and exported as encoded properties files. These formed the basis for IRIS to be in the 6 official languages.

User Interface

Apart from look and feel customization to match the existing www.who.int site, DSpace also needed further enhancements to accommodate a good rendering of right to left for the Arabic and to also render this when metadata had a mixture of RTL and LTR. Some aspects of making sure that the other character sets for the 6 languages would render well required some customization or fixing as well.