Federated content seach use case at UPF

This is a real use case we addressed last month:
People from the *Institut d'Estudis Catalans* (a sort of local royal academy) asked us for help. They need to identify all 'books' translated from Latin or Greek into Catalan.

As you can see the use case does not involve 'pure content' search (in the sense that it does not imply corpus querying) and it is not very exciting. However it is a real request and it is interesting as far as metadata searching is concerned and demonstrates that current OAI-PMH harvesting cannot cope with such a 'trivial' query.

We searched big repositories such as LOC (SRU) and The European Library (Open Search) and we failed. Big repositories harvest metadata in DC format. This means that fine grained descriptions in remote repositories (in MARC format, for example) are collapsed into DC descriptions. This makes it hard to search for things such as 'translator', 'source language' or 'target language' as this information is lost or embedded in more general fields.

We tried to use content search in metadata records using queries such as:

search for 'creators' with 'trad.' (Sometimes translators are included in 'creator' field):
http://hispana.mcu.es/i18n/sru/sru.cmd?query=**dc.creator+any+%22trad.%22**

search for titles with 'trad.' (Sometimes translators are included in 'title'):
http://hispana.mcu.es/i18n/sru/sru.cmd?query=**dc.title+any+%22trad.%22**

The idea was to

1. Get Catalan 'books' with 'translators'
2. For each 'book', identify the author and get his/her works
3. perform language guesser on these works to identify 'original' language of the author

We fail at (1).

In our case, once again 'data collection' was the real bottleneck: researchers want to perform experiments on very specific data sets. We are ready to perform nice experiments but it is hard to find the data they want.