# comments on the search function for AGATE

Jochen Schirrwagen, 16.01.2017

* general comment: a lot of similarities with OpenAIRE, both from the point of a human infrastructure (e.g. helpdesk) but also technical infrastructure

## metadata guidelines / format

* if you plan to release guidelines
  * define mandatory, recommended, optional metadata properties
  * not all data providers will be able to deliver all metadata, make clear what is required as a minimum
  * make use of controlled vocabularies whenever possible otherwise you'll have a big issue with metadata normalization and data loss
  * also it makes indexing and search easier
  * suggest the use of (subject-specific) thesaury if applicable (either already in the metadata records or in your search interface/portal)
  * compare your requirements with existing metadata guidelines and application profiles
  * avoid to reinvent the wheel, rather decide to re-use and extend an existing standard
  * also it will make life of repository managers easier if they do not need to support different requirements from different service providers / aggregators
  * guidelines need good documentation complemented by support and training
  * implementation of guidelines might be challenging if diverse (repository) platforms must be supported
  * the lack of granularity in (simple) dublin core is a big issue, choosing a more granular (and established) format is an option while a format that is compatible with linked data principles might be a good investment for the near future

## metadata aggregation

* expose metadata as a hub is a very good idea since other aggregators can benefit from. E.g. Openaire already aggregates from some national and regional nodes beside from individual repositories and journals.

## metadata curation effort

* do not underestimate the effort for metadata curation, healing, normalization even if guidelines are in place
* consider the effort needed for the communication with repository managers (in both directions); feedback loop
* metadata healing might be a more sophisticated but realistic task:
  * few infrastructures like the JISC-Repository Junction Broker, the OpenAIRE Notification Broker (currently in prototype status) allow for subscription and notification of events (repo A has metadata reference only but repo B has the same metadata but with open access fulltext; record from repo C has an incomplete list of authors but the record from repo D about the same publication is complete)

## workflow

* for each step in the workflow (according to the conceptual diagram of the search function) make roles and responsibilities clear
* e.g. where does the linkage between entities (publications, research data, software, project) takes place, on the level of the repository or in AGATE?
* what features are provided to end-users, how can they interact with the system? can they enrich existing information?

## tools

* as a minimum I would consider an aggregation environment (supports but is not limited on OAIP-PMH) with normalization features, Solr for indexing, lightweight portal solution (maybe separate from a CMS?)

## api
OAI-PMH of course widely established but more and more critized to be out of date; be prepared for alternativ solutions in the near future ;-)  
* REST-API, FTP  
* web-crawling, web-scraping (e.g. from project homepages) as custom and workaround solutions