Repository content – lost in discovery?

Open Repositories, Helsinki, June 12, 2014
Jyrki Ilva (jyrki.ilva@helsinki.fi)
Discovery systems

- The basic idea: giving the patrons a single search interface into content (both local and non-local) provided by the library.
Repositories and discovery

- How does repository content show up in discovery systems, or does it show up at all?
Discovery in Finland

- A national discovery system, Finna (www.finna.fi) for Finnish archives, libraries and museums
- Part of the National Digital Library Project (KDK, www.kdk.fi/en) funded by the Ministry of Education and Culture
Development of Finna

- Developed by the National Library of Finland
- Based on VuFind and a number of other tools
- Lots of good work done, see https://github.com/KDK-Alli
- Provides both a national interface for publicly available content and also separate instances for individual organizations
Using Finna on an organizational level (1)

- Helsinki City Museum, https://hkm.finna.fi/
Using Finna on an organizational level (2)

- University of Jyväskylä, https://jyu.finna.fi/jykdok/
- Designed to replace the university level OPAC and meta-search tool

![Search results for scientific publishing on JykDok](image)

- **Productivity in academia: scientific publishing at Norwegian universities**, Book, by Kyvik, Stian, Publisher: 1991, Library of the Institute for Educational Research (Building Opisni), Loan period 38 days, More options (Open in a new window)
Repository content in Discovery?

- Finland currently has about a dozen repositories, with 50+ organizations and 200.000+ OA full-text publications
- VuFind/Finna can harvest metadata from repositories
- However, it is not always clear how to deal with the repository content in discovery context
Discovery: Conceptual issues

- Finna has two separate indices – a local index and a central index
  - The local index contains harvested metadata for the content which may or may not have been previously catalogued into the library OPAC
  - The central index contains metadata for the licensed content acquired by the library
- Should the worldwide repository content – and other non-local OA content – be available either in the local index or in the central index, or both?
Repositories in a local index

- What should there actually be in the local index?
  - There seems to be a tendency to restrict the local index strictly to the local content
  - For example, only theses produced within the organization, not from elsewhere
  - There are fears that the local content will get lost in the search results, if there is too much outside stuff available
- So, we may end up having the content of the local repository harvested, but not the repository content from elsewhere
Repositories in a central index?

- There are several big commercial vendors providing a central index for licensed content
  - Can we trust them to include open access repository content from all around the world in their product?
  - Currently the short answer is no: to a large extent, the vendors seem to harvest repository content only from their own customer organizations
We are not making it easy, either

- The repositories contain very diverse stuff – should all of this be harvested?
- The repository metadata is far from uniform and has quality issues
- These problems have been well-known for a long time… but we haven’t been able to solve them
  - OAIster wasn’t much of a success
  - BASE ([http://www.base-search.net/about/en/](http://www.base-search.net/about/en/)) is much better, but also provides excellent illustration for many of the problematic issues
- Discovery systems like Finna must deal with the same issues, if we hope them to include the repository metadata as well
Challenges: Metadata formats

- Dublin Core
  - No clearly defined or easy-to-understand guidance on how to deal with metadata that goes beyond the basic 15 fields
  - Leads to different local interpretations, ad hoc solutions
  - The development of DC itself seems to have just about stalled
Challenges: Google is making us lazy?

- Despite these problems, we have been doing quite well, thanks to Google.
- Most of us are getting most of our traffic via Google ... which doesn’t really rely on the quality of our metadata.
Challenges: Fragmented community

- Unfortunately, the repository community is fairly diverse and fragmented
- There are valiant efforts to solve metadata issues
  - Some may be platform-based
  - Some may aim for coherence on a geographical area
- But it’s hard to come up with a global solution that would be adopted by everybody
Should we care? What should we do?

- We may continue to rely on Google even in the future…
  - But what if it really turns evil at some point?
- We should work with the discovery people to make sure that our content is included in a sensible way
- Even if it’s not going to be easy, we should do support the efforts for metadata co-ordination
How to fix it for Finna?

- Of course, we just cannot wait for the issues to be solved on a global level (here we go again…)
- We must figure out how much normalization can be done after the metadata is already harvested from local systems
- And we must also figure out what can be done to improve the coherence of metadata in local repositories, at least on a national level
Thanks!

jyrki.ilva@helsinki.fi