Data Searchery

Preliminary Analysis of Data Sources Interlinking

Paolo Manghi and Andrea Mannocci

Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
name.surname@isti.cnr.it

Abstract. The novel e-Science's data-centric paradigm has proved that interlinking publications and research data objects coming from different realms and data sources (e.g. publication repositories, data repositories) makes dissemination, re-use, and validation of research activities more effective. Scholarly Communication Infrastructures are advocated for bridging such data sources, by offering tools for identification, creation, and navigation of relationships. Since realization and maintenance of such infrastructures is expensive, in this demo we propose a lightweight approach for "preliminary analysis of data source interlinking" to help practitioners at evaluating whether and to what extent realizing them can be effective. We present Data Searchery, a configurable tool enabling users to easily plug-in data sources from different realms with the purpose of cross-relating their objects, be them publications or research data, by identifying relationships between their metadata descriptions.

Keywords: Interoperability, Interlinking, Research Data, Publications.

1 Introduction

The Research Digital Libraries (RDLs) ecosystem is ever growing since creating and publishing a proprietary *publication repository* is essentially mandatory for any institution striving to gain a modicum of visibility and relevance. In addition, the advent of e-Science and data-intensive research [1] has fired a similar trend for research data. *Data repositories* for persisting and publishing research data are becoming common in many scientific communities [2].

In such a scenario, being able to correctly infer relationships among objects belonging to different domains, i.e. publication-publication, publication-data and data-data interlinking, becomes crucial in order to: (*i*) foster multi-disciplinarity by looking at adherences among distinct disciplines; (*ii*) enable a better review, reproduction and re-use of research activities [3]. Identifying which *data sources* are worth bridging in such a plethora of publication and data repositories from different scientific domains is not an easy task. Nonetheless, understanding which kind of relationships can be inferred across objects of different data sources is yet another challenge. Finally, interoperability issues generally arise, since access protocols, metadata formats and object models are likely to differ for different data sources, due to technological and scientific domain peculiarities. Scientific

T. Aalberg et al. (Eds.): TPDL 2013, LNCS 8092, pp. 458–461, 2013.

communities cope with these needs by realizing *Scholarly Communication Infrastructures* (SCIs). These provide tools and services to aggregate objects coming from different data sources and realms and enable both humans and machinery to interconnect such objects by identifying relationships via user-interfaces or advanced inference-by-mining algorithms (e.g. OpenAIRE [4]).

Since requirements change both from case to case and over time, SCIs have to specifically address ever changing requirements and therefore must be planned and designed very carefully. Additionally, once deployed they have to undergo a continuous and expensive process of extension, optimization and maintenance. Thus, their cost in terms of time and skills tends to be generally high and sometime prohibitive for the smallest communities. For such reasons, planning the realization of a SCI would benefit from tools permitting a lightweight preliminary analysis of data source interlinking possibilities. In this demo we present the prototype of Data Searchery, a tool aiding practitioners willing to realize SCIs to evaluate whether and to what extent the intended SCI could be effective for the community. The tool offers user interfaces to run and cross advanced metadata searches over data sources and therefore identify relationships between their objects. Data Searchery may also be useful for scientific user communities willing to identify relationships between objects between data sources not yet aggregated and interconnected in the context of a SCI.

2 Data Searchery Overview

Data Searchery enables users to surf and (best-effort) correlate metadata coming from two different data sources, here referenced as *origin* and *target* data sources, which may contain publication or research data objects. A data source can be either a sole repository serving a single institution or an aggregative data infrastructure serving a research community; e.g. DataCite[1] for research data. The approach relies on the metadata descriptions of the data source objects. The more metadata are accurate and thorough, the more the recall of the approach tends to be accurate, but in general, poor or incomplete metadata content does not necessarily invalidate the generic reasoning behind the approach. As shown in Fig.1, the prototype guides the user in a two-step process: (*i*) querying an origin data source of preference; (*ii*) starting from one record of interest returned by the first query, drafting a second query on a target data source of preference. The user drafts the first query by typing keywords and (optionally) by narrowing down to one or more collections made available by the origin data source. The second query is drafted by automatically extracting keywords from one metadata record of interest, chosen among the ones returned by the first query. Keywords are extracted by applying one or more *extraction filters* selected by the user from a list; as in the first step, the user can narrow the query down to one or more collections exposed by the target data source.

Harvesting, storing and cleaning metadata falls out the scope of Data Searchery which on the contrary relies solely on "live-queries" over remote data sources;

[1] The DataCite Initiative, `http://datacite.org/`

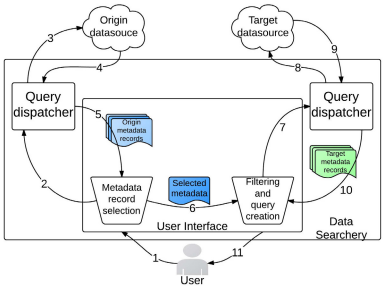

Fig. 1. Data Searchery interlinking session

no standard protocol such as OAI-PMH or OAI-ORE is thus involved. The only requirement is the availability of a search API at the data source (e.g. Solr API). Relevant information is extracted from results and rendered onto the screen as the search API provides it (e.g. ordered by field relevance in the case of Solr index) without any additional re-ordering or proximity-scoring policy. Data Searchery is designed to be a general-purpose tool handling the abstractions: *data source*, i.e. an API capable of responding to keyword queries and collection queries, *data source sub-collections*, i.e. an API returning the list of possible data source collections, and *extraction filters*, i.e. functions inferring a set of keywords from an input record (e.g. identifying and extracting given ontology terms from the *abstract* field). The user interface adapts itself seamlessly to additions, changes, or removal of such abstractions. Data Searchery is open source and developed in Java, thus developers can easily add instances of data sources and extraction filters as new implementations of corresponding Java classes.

A running instance of the Data Searchery prototype with some built-in data sources and extraction filters can be found here: http://datasearchery-prototype.research-infrastructures.eu/datasearchery#/search.

3 The Demonstration

During the demo, users will first be driven through explanatory and meaningful interlinking sessions, then given free access to the tool. Fig.2 shows the tool's UI, the search panel on the left-side. The user has selected DataCite as original data source and the OpenAIRE infrastructure [4] as target data source, choosing from a drop down menu of available data sources. Her/his first search is on DataCite with the keywords *"Abyssal brown algae"*. The search results are fetched and rendered on the left side of a two-column layout. then the user is interested in finding cross-links between the dataset *"n-Alkanes and phenols in bottom. . . "* with publications in OpenAIRE. To this aim, she/he selects both keywords and organisms extraction filters from the record drop-down menu[2] and fires the search. The results are displayed on the right-side column in the screen and dynamically adapt to changes in the search parameters.

[2] WhatIzIt - EBI, `http://www.ebi.ac.uk/webservices/whatizit/`

It must be noticed that the prototype relies entirely on dynamic interaction with data sources, hence the same interlinking session, run at different times, may return different number of hits or results ordering depending on the changes to the original content.

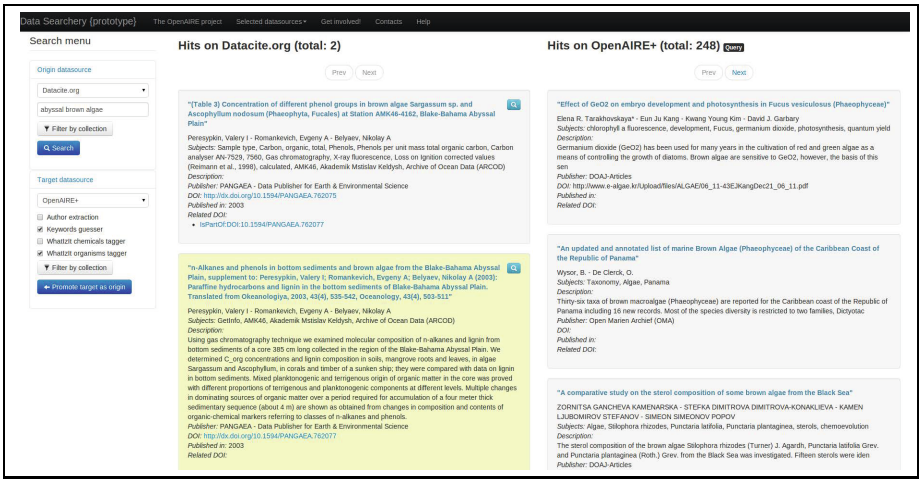

Fig. 2. Screenshot of Data Searchery web app

4 Conclusions and Future Work

The ability to correlate either data or publications hosted in different data sources and realms is becoming a key aspect in modern scholarly communication. SCIs advocate this new trend, but their realization and maintenance raises serious sustainability issues. Data Searchery is a configurable tool allowing for lightweight and preliminary evaluation of the existence of meaningful links between objects from different data sources. Currently the tool is being extended with functionalities to elaborate extensive statistical reports on the overall degree of correlation between two data sources w.r.t. a set of user queries (e.g. a pool of authors and/or keywords).

References

1. Gray, J.: A transformed scientific method. In: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research (2009)
2. Callaghan, S., Donegan, S.: Making data a first class scientific output: Data citation and publication by NERC's environmental data centres. International Journal of Digital Curation (2012)
3. Reilly, S., Schallier, W., Schrimpf, S., Smit, E., Wilkinson, M.: Report on integration of data and publications. ODE Opportunities for Data Exchange
4. Manghi, P., Bolikowski, L., Manola, N., Shirrwagen, J., Smith, T.: Openaireplus: the european scholarly communication data infrastructure. D-Lib Magazine 18(9-10) (September-October 2012)