Preliminary Analysis of Data Sources Interlinking

Data Searchery: a case study

Andrea Mannocci and Paolo Manghi

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 (SCIs) are advocated for bridging such data sources by offering an overlay of services for identification, creation, and navigation of relationships among objects of different nature. Since realization and maintenance of such infrastructures is in general very cost-consuming, in this paper 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 delivering a service for relating objects across data sources, be them publications or research data, by identifying relationships between their metadata descriptions in real-time.

Keywords: Interoperability, Interlinking, Research Data, Publications, Metadata, Inference

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. However, despite printed papers and their digital shadows still are and will remain the principal carriers for research outcome dissemination, they will never be *per se* adequate to embed huge datasets, software, code listings, workflows, media files and any other sidecar-information which simply cannot fit into the traditional publishing model [1, 2]. This drawback is aggravated by the advent of e-Science and data-intensive research [3] which turned research data into a first class citizen in research processes and output [4, 5]. *Data repositories* for publishing, curating and persisting research data are becoming more and more common in many scientific communities [5] as well as *data papers*, peer-reviewed articles which are intended as a thorough description of a dataset (or a group of datasets) and the way they have been produced [6].

For such reasons, the availability of overlay services capable of bridging data and publication worlds is getting more and more appealing for research communities [4]. In particular, inferring 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 [7]; (*ii*) enable a better review, reproduction and re-use of research activities [4, 1]. However, identifying which domains are worth bridging in such a plethora of publication and data repositories is not an easy task. In addition, understanding which kind of relationships can be inferred across objects of different domains is yet another challenge. Finally, interoperability issues generally arise, since access protocols, metadata formats and object models are likely to differ for different repositories, due to technological and scientific domain peculiarities.

In order to solve such issues, scientific communities lately started realizing *Scholarly Communication Infrastructures* (SCIs) [8], providing tools and services to aggregate object files and metadata coming from different data sources (e.g. repositories or other SCIs) and scientific realms and enable both humans and machinery to interlink such objects by identifying relationships via user-interfaces or advanced inference-by-mining algorithms (e.g. OpenAIRE and OpenAIREplus [9, 10], Mendeley[1], ORCID[2], Utopia Documents [11]).

Since requirements differ 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. In particular, the majority of the total costs are due to data source aggregation efforts. For such reasons, planning the realization of a SCI would benefit from tools assessing metadata interlinking capabilities among data sources in a preliminary lightweight fashion without the need of any prior object file and/or metadata collection and processing.

In this paper we present Data Searchery, a tool conceived and developed within the efforts of the two FP7 projects OpenAIREplus [9, 10] and iCORDI[3]. Data Searchery is intended to help practitioners willing to realize new SCIs or extend existing SCis with functionalities to interlink objects across data sources at evaluating whether and to what extent such an investment would be effective for the community. To this aim, the tool offers an intuitive interface enabling end-users to identify meaningful relationships between objects of given data sources via advanced cross-search mechanisms over their metadata descriptions. Data Searchery is designed to facilitate the integration of new data sources into the

[1] Mendeley, http://www.mendeley.com/

[2] ORCID, http://orcid.org/

[3] The Research Data Alliance, http://europe.rd-alliance.org/

framework as well as flexibly integrate key-word extraction filters to specialise cross-metadata search to given scientific fields.

Outline. The remainder of the paper is organized as follows. Section 2 provides functional requirements and an high-level overview of Data Searchery architecture and design. Section 3 describes our real-life experience gathered through the use of Data Searchery within the OpenAIREplus project efforts. Section 4 describes related work. Finally Section 5 makes final remarks and paves the way for future enhancements of the method.

2 Data Searchery Overview

Because of the dramatic impact the aggregation phase has over global cost of SCIs development, we advocate here the convenience for a tool able to preliminary asses the feasibility and effectiveness of efforts in the aggregation phase. Given the high availability of repositories over the Internet, we want to let the user search for and play with metadata there exposed and be able to surf and (best-effort) *correlate on-the-fly* metadata present in two different *data sources*, here referenced as *origin* and *target* data sources, which may contain both publications and research data objects. For the sake of flexibility and effectiveness, such a tool should also offer some mechanism to leverage in order to *filter and extract* relevant information from metadata records for driving new searches and inferring relationships.

Such features are traditionally offered by the service overlay exposed by SCIs, alas available only at the end of all design and implementation efforts. When not available, the user has to face with the discovery of repositories, the individual querying process for each one of them and the manual comparison of returned results scattered across multiple search portals. By taking advantage of a lightweight mechanisms for preliminary exploration of interlinking capabilities among data sources, the advantage is twofold. On one side it would cut unneeded costs by preventing ineffective tasks to take place during SCIs development, on the other side traditional users usually performing multiple queries juggling between several search pages would benefit of a single-paged UI providing all the tools needed for running guided and meaningful cross-queries.

2.1 Scenario and Functional Requirements

In the following, we refer to a *data source* as a entry-point publicly queryable from the Internet able to offer objects (i.e. *metadata records*) optionally organized into sets, hereafter mentioned as *collections*, against whom it is possible to narrow queries down. A data source can be either a publication repository, or a dataset repository, or a repository aggregator (e.g. DataCite[4], NARCIS[5]), or

[4] The DataCite Initiative, http://datacite.org
[5] NARCIS, http://www.narcis.nl/

a SCI (e.g. OpenAIRE[6], DRIVER[7], Google Scholar[8]). Each data source adopts its own metadata schema (either proprietary or standardized), but in general we assume a metadata format can be considered as a flat list of $< field, type >$ couples.

Given two data sources, common sense and experience would suggest the user to feed the same query to both (possibly via web portals) and optionally refine the second query with additional information coming from a record of interest out of the results of the first query. This *refinement process* has been modeled in Data Searchery with the concept of the *extraction filter*. Hereafter, we refer to extraction filter as an abstraction capable of automatic extraction of textual and semantic features given one metadata record of preference; these additional information can be used in subsequent queries in order to refine results and infer more precise relationships among objects. For example, semantically relevant keywords can be extracted from the title of the selected metadata record and can be used for running a second query on the target data source.

The Data Searchery approach relies entirely on metadata descriptions of objects provided by data sources, hence the more metadata are accurate and thorough, the more the recall of the tool tends to be accurate. However in general, poor or incomplete metadata content does not necessarily invalidate the generic reasoning behind the approach.

Data Searchery is a tool aiming to satisfy the following typical use-cases. The former describes what anticipated so far and reflects the point of view of the final user enjoying a service for cross-searching data sources in couples availing of a set of configurable extraction filters in order to gather contextual information out of metadata. The latter describes the developer perspective in customizing a Data Searchery instance by adding additional data sources and extraction filters in a flexible and easy way.

Data Searchery final user. As shown in Fig.1, the prototype guides the user in a two-step process: (*i*) querying an origin data source of preference; (*ii*) given one record of interest returned by the first query, crafting 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 tool-crafted 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 present on the target data source.

The solution has to be as generic and flexible as possible and make the user able to navigate metadata starting either from research data or scientific publications indifferently. Furthermore, it should be possible to "promote" a target data source into an origin data source for subsequent queries; in this way,

[6] The OpenAIRE project, http://www.openaire.eu/en

[7] The DRIVER repository, http://www.driver-repository.eu/

[8] Google Scholar, http://scholar.google.com

the aforementioned scenario can be re-iterated and lead to a new interlinking step targeting a third different data source and so on.

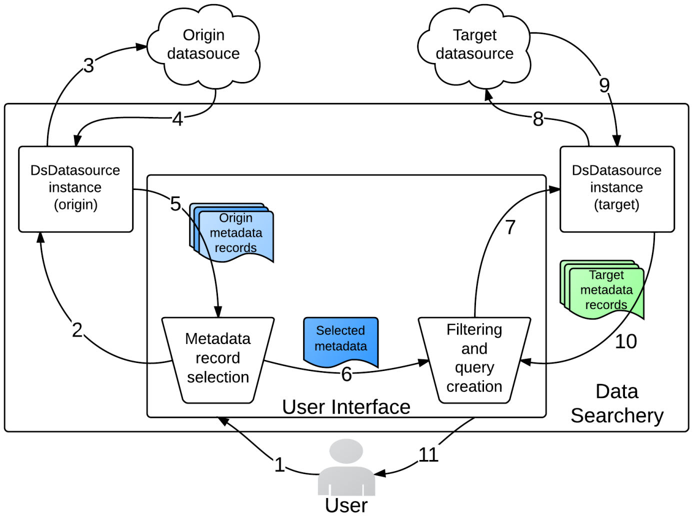


Fig. 1. Data Searchery interlinking session

Data Searchery developer. A programmer in charge of developing a customized instance of Data Searchery needs to plug-in both data sources and extraction filters in an easy and flexible way through minimal changes (hopefully only adding new implementation rather than modifying pre-existing one). Finally, Data Searchery user interface should be to some extent oblivious to the addition and modification of such abstractions and adapts itself seamlessly to such changes.

2.2 Architectural and Design Overview

Data Searchery is open source and developed in Java; it interfaces real-world data source via the abstraction named `DsDatasource` which is in charge of fetching and map native metadata records to Data Searchery format. Similarly, the refinement process extracting context information out of metadata records is performed via functions called `DsExtractionFilter`. Developers can easily plug-in new instances of data sources and extraction filters as new implementations of the corresponding Java classes briefly described below.

DS Data Source. The `DsDatasource` abstraction requires implementing two API methods capable of (*i*) responding to *keyword queries* and returning a page of metadata records (i.e. `getMetadataPage()::List<DsMetadataRecord>`) and (*ii*) *collection queries* returning a list of collections available at the data source (i.e. `getCollections()::List<DsCollection>`).

The first API method is responsible for (*i*) querying a data source and wait for results, and (*ii*) mapping returned native metadata records into Data Searchery metadata records (i.e. `DsMetadataRecord`). We assume the minimal requirement for a data source to be compliant with Data Searchery is the availability of a search API (e.g. Solr[9] or Elasticsearch[10] APIs) accepting textual queries and returning paginated *XML-formatted results* (i.e. metadata records). It must be noticed that harvesting, storing and cleaning metadata fall out the scope of Data Searchery which on the contrary relies solely on "live-queries" over remote data sources; no standard protocol such as OAI-PMH or OAI-ORE is thus involved. Since each data source generally adopts a different metadata schema, for the sake of simplicity we designed a simple model class `DsMetadataRecord` onto which mirror relevant information for interlinking purposes contained inside original metadata fields. For each native metadata record returned by a queried data source, a `DsMetadataRecord` instance is created (stored in memory with no caching mechanism) and populated with information from the original metadata. A `DsDatasource` maps native records as the search API provides them (e.g. ordered by field relevance in the case of Solr index) without applying any additional re-ordering or scoring policy. More precisely, for each native metadata record it is possible to map authors, title, abstract/description, keywords/subject, publisher and native repository in our model. Since some metadata format might carry a list of *related DOIs* (both within the same data source and external), saving this identifiers is useful too as it already solves partially the problem statement by providing a set of surely related items. When available, publication date is mapped too for the sake of completeness, so that time-aware interlinking queries can be performed. The simple model described here can be extended as needed with due modifications to `DsMetadataRecord` class; the changes introduced are reflected seamlessly onto the web UI without any additional coding effort.

The second API method provides a list of collections optionally exposed by a data source. Such list of available collections could be either hard-coded (hence statically loaded), or dynamically retrieved from the Internet by several means (e.g. from a web resource or from the filesystem). According to the data sources examined so far, the idea of collection has been implemented by indexing (for each record) a special field containing to the name of the collection of provenance. Narrowing queries against this field gives the impression of querying a logical index reserved to the collection.

An an example, regarding the DataCite data source, in our prototype we created a new `DataciteDS.java` class deriving `DsDatasource.java` which maps native records from Datacite metadata format into `DSMetadataRecord`, and retrieves the collection list by parsing a plain text web page hosted at DataCite facilities. Similarly, for the Driver project [12, 13], we implemented another class (`DriverDS.java`) derived from `DsDatasource.java` which maps from Driver

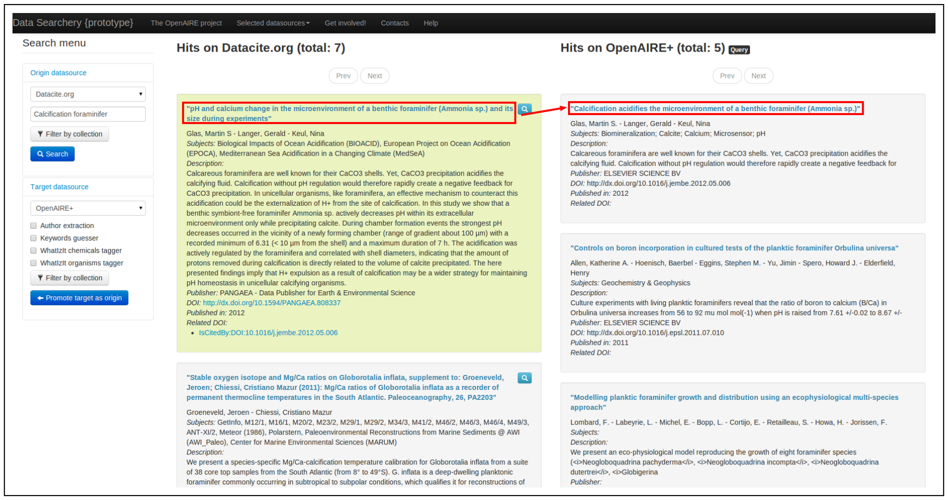
[9] Apache Solr, http://lucene.apache.org/solr/

[10] Elasticsearch, http://www.elasticsearch.org/

metadata format and fetches the collection list by parsing an XML-formatted page returned by a query performed over D-Net[11] Information Service, the data infrastructure on top of which the project has been realized.

DS Extraction Filter. Extraction filters are instantiated by extending `DsExtractionFilter` abstraction which requires implementing the function `processMetadataRecord()::List<String>` and returns a set of inferred keywords given an input metadata record. In particular, an extraction filter operates over a set of fields of a `DsMetadataRecord` object trying to identify and extract given ontology terms or semantically meaningful keywords. The logic for such extraction function can be implemented "in-house" as local implementation or can be demanded to an external service. For example, in our prototype the two text taggers invoke an external service made available by EBI [12]. On the contrary, all the others extraction filters implemented in the prototype rely on local implementation. Extraction filters can be selectively associated to the data sources for which they are needed. For example, a "keyword finder" extraction filter which is looking for keywords semantically related to the world of biodiversity should not be available for data sources offering technical science metadata, or at least it would not make sense in principle.

A running instance of the Data Searchery prototype, in Fig. 2, with sample built-in data sources (DataCite, DRIVER, OpenAIREplus) and extraction filters (author extractor, keyword extractor, EBI text taggers) can be found online at: http://goo.gl/pI3VR.



Datacite.org ▾

Calcification foraminifer

OpenAIRE+ ▾

□ Author extraction
□ Keywords guesser
□ WhatIzIt chemicals tagger
□ WhatIzIt organisms tagger

Hits on Datacite.org (total: 7)

"pH and calcium change in the microenvironment of a benthic foraminifer (Ammonia sp.) and its size during experiments"

Glas, Martin S - Langer, Gerald - Keul, Nina
Subjects: Biological Impacts of Ocean Acidification (BIOACID), European Project on Ocean Acidification (EPOCA), Mediterranean Sea Acidification in a Changing Climate (MedSeA)
Description:
Calcareous foraminifera are well known for their CaCO3 shells. Yet, CaCO3 precipitation acidifies the calcifying fluid. Calcification without pH regulation would therefore rapidly create a negative feedback for CaCO3 precipitation. In unicellular organisms, like foraminifera, an effective mechanism to counteract this acidification could be the externalization of H+ from the site of calcification. In this study we show that a benthic symbiont-free foraminifer Ammonia sp. actively decreases pH within its extracellular microenvironment only while precipitating calcite. During chamber formation events the strongest pH decreases occurred in the vicinity of a newly forming chamber (range of gradient about 100 µm) with a recorded minimum of 6.31 (< 10 µm from the shell) and a maximum duration of 7 h. The acidification was actively regulated by the foraminifera and correlated with shell diameters, indicating that the amount of protons removed during calcification is directly related to the volume of calcite precipitated. The here presented findings imply that H+ expulsion as a result of calcification may be a wider strategy for maintaining pH homeostasis in unicellular calcifying organisms.
Publisher: PANGAEA - Data Publisher for Earth & Environmental Science
DOI: http://dx.doi.org/10.1594/PANGAEA.808337
Published in: 2012
Related DOI:

- IsCitedBy:DOI:10.1016/j.jembe.2012.05.006

"Stable oxygen isotope and Mg/Ca ratios on Globorotalia inflata, supplement to: Groeneveld, Jeroen; Chiessi, Cristiano Mazur (2011): Mg/Ca ratios of Globorotalia inflata as a recorder of permanent thermocline temperatures in the South Atlantic. Paleoceanography, 26, PA2203"

Groeneveld, Jeroen - Chiessi, Cristiano Mazur
Subjects: GetInfo, M12/1, M16/1, M20/2, M23/2, M29/1, M29/2, M34/3, M41/2, M46/2, M46/3, M46/4, M49/3, ANT-XI/2, Meteor (1986), Polarstern, Paleoenvironmental Reconstructions from Marine Sediments @ AWI (AWI_Paleo), Center for Marine Environmental Sciences (MARUM)
Description:
We present a species-specific Mg/Ca-calcification temperature calibration for Globorotalia inflata from a suite of 38 core top samples from the South Atlantic (from 8° to 49°S). G. inflata is a deep-dwelling planktonic foraminifer commonly occurring in subtropical to subpolar conditions, which qualifies it for reconstructions of

Hits on OpenAIRE+ (total: 5)

"Calcification acidifies the microenvironment of a benthic foraminifer (Ammonia sp.)"

Glas, Martin S. - Langer, Gerald - Keul, Nina
Subjects: Biomineralization; Calcite; Calcium; Microsensor; pH
Description:
Calcareous foraminifera are well known for their CaCO3 shells. Yet, CaCO3 precipitation acidifies the calcifying fluid. Calcification without pH regulation would therefore rapidly create a negative feedback for
Publisher: ELSEVIER SCIENCE BV
DOI: http://dx.doi.org/10.1016/j.jembe.2012.05.006
Published in: 2012
Related DOI:

"Controls on boron incorporation in cultured tests of the planktic foraminifer Orbulina universa"

Allen, Katherine A. - Hoenisch, Baerbel - Eggins, Stephen M. - Yu, Jimin - Spero, Howard J. - Elderfield, Henry
Subjects: Geochemistry & Geophysics
Description:
Culture experiments with living planktic foraminifers reveal that the ratio of boron to calcium (B/Ca) in Orbulina universa increases from 56 to 92 mu mol mol(-1) when pH is raised from 7.61 +/-0.02 to 8.67 +/-
Publisher: ELSEVIER SCIENCE BV
DOI: http://dx.doi.org/10.1016/j.epsl.2011.07.010
Published in: 2011
Related DOI:

"Modelling planktic foraminifer growth and distribution using an ecophysiological multi-species approach"

Lombard, F. - Labeyrie, L. - Michel, E. - Bopp, L. - Cortijo, E. - Retailleau, S. - Howa, H. - Jorissen, F.
Subjects:
Description:
We present an eco-physiological model reproducing the growth of eight foraminifer species (<i>Neogloboquadrina pachyderma</i>, <i>Neogloboquadrina incompta</i>, <i>Neogloboquadrina dutertrei</i>, <i>Globigerina
Publisher:

Fig. 2. Screenshot of Data Searchery web app

[11] D-Net Software Toolkit, http://www.d-net.research-infrastructures.eu

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

3 OpenAIREplus: an use case

The OpenAIREplus project (2^{nd} generation of the former project OpenAIRE) [9, 10] deals with the delivery and operation of the OpenAIRE scholarly communication infrastructure, currently integrating about 400 OpenAIRE-validated repositories across Europe (from a total of 2,500 OpenDOAR publication repositories and 250 re3data.org data repositories). The SCI also offers inference-by-mining algorithms and search mechanisms over 8,000,000 publications and 1,500,000 datasets.

Being more accurate, in OpenAIREplus inference is performed over harmonized metadata present in the information space (i.e. right after aggregation and cleaning phases) and mainly focuses on (*i*) identifying project-publication relationships (funding sources), (*ii*) publication-datasets correlation, and (*iii*) publications-publications inference w.r.t. citation, similarity, duplicates [9].

Data Searchery has been tested in a few use cases within the efforts of the OpenAIREplus project in order to preliminary explore adherences between couples of selected data sources and examine metadata prior to aggregation phase. For example, through the use of Data Searchery, we were able to perform several positive interlinking sessions confirming that all the premises for an effective interlinking among dataset objects from DataCite repositories and publication objects from OpenDOAR repositories were righteous in the first place. A sample session is one searching for the keywords *"Calcification foraminifer"* on the DataCite repository as origin data source. The query returns 7 metadata records, the first of which describes a dataset titled *"pH and calcium change in the microenvironment of a benthic foraminifer (Ammonia sp.) and its size during experiments"* by Glas *et al.*, published by PANGAEA[13] repository, one of the partner joining the DataCite initiative. By (*i*) searching for related items having selected OpenAIREplus as target data source and (*ii*) having activated keywords and authors extraction filters, we discover a publication by the same authors titled *"Calcification acidifies the microenvironment of a benthic foraminifer (Ammonia sp.)"* which refers the same dataset found at the previous step and that is stored at "Web of Science" repository. As a result of further positive sessions, the analysis suggested that publications in WoS and PANGAEA are likely related by discipline bindings, hence that investigating on more advanced infrastructure interlinking services capable of identifying and persisting such links are likely worth the effort.

4 Related Work

To the best our knowledge, literature on cross metadata search of web data sources with the purpose of interlinking data and publication is lacking. On the contrary, literature on Linked Open Data[14] (LOD) and RDF interlinking is copious. In [15], Wölger *et al.* draw some indicators for the classification of

[13] PANGAEA - Data Publisher for Earth & Environmental Science, http://www.pangaea.de

methods and frameworks for Linked Open Data and RDF datasets interlinking. As the classification seems reasonable and sufficiently generic to contextualize inference mechanisms for metadata interlinking, we refer to their work which is focused on the following key indicators:

degree of automation whether the tool needs human intervention or not and to which extent;

human contribution the type of interaction an user has to provide in order to be able to use effectively the tool;

domain whether the method is bound to a specific domain or domain independent;

matching techniques the mechanisms the approach leverages in order to match data;

ontologies whether ontologies are taken into account or not;

input what and in which format has to be fed to the tool;

output what and in which format the tool returns;

post-processing post-processing capabilities offered by the tool;

data access protocol or method used for accessing data.

Let us extend this classification by adding a couple of novel indicators to the list: *real-timeliness* and *volatility*. A tool is classified as a *real-time* tool if it tries to infer links on-the-fly and does not require any *a priori* knowledge or processing of datasets; a tool is classified as *deferred* otherwise, since it fetches required metadata and then processes it looking for inference (e.g. several methods analyzed in [15] belong to this category). A tool is declared *volatile* if persisting inferred relationships falls out the scope of the approach (i.e. the tool is only exploratory); the tool is said to be *persistent* otherwise, since links are someway persisted for future navigation.

According to the proposed extension, the Data Searchery prototype can be classified as follows:

real-timeliness fully real-time. No pre-processing needed;

degree of automation manual (semi-automatic mode is a work in progress);

human contribution selected data sources, keywords and search parameters;

domain domain independent;

matching techniques string matching, word relation matching (synonyms, taxonomic similarities);

ontologies pluggable as extraction filters;

input one XML-formatted metadata (out of results returned from the query performed on the origin data source) plus an extraction filters configuration;

volatility volatile;

output a page of XML-formatted metadata returned from the target data source;

post-processing n/a;

data access searchable API (e.g. SOLR index API, elasticsearch API).

In [16], Nikolaidou *et al.* describe a meta-search service for gathering knowledge about music genres, bands and discographies by combining metadata results coming from three different data sources (MusicBrainz, Last.fm, and Discogs). In their approach, the three data sources are queried separately and results retrieved by one become a possible input for a more specialized search in the other two; results are combined at real-time as the user proceeds with his/her interrogation and nothing is persisted on disk.

5 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 as it fosters findability, review and reuse of previous work and promotes multi-disciplinarity and idea dissemination across communities. SCIs tackle this new trend by offering a service overlay and tools tackling this need, however their realization and maintenance raise serious sustainability issues and makes them seldom viable for the smallest communities.

Being able to infer on-the-fly relationships between objects of different nature and belonging to different contexts without any prior metadata harvesting and processing would be an appealing perspective. Data Searchery is a configurable tool allowing for lightweight and preliminary evaluation of the existence of meaningful links between objects from different data sources. The tool offers an intuitive web interface allowing users to surf and correlate on-the-fly metadata across two different data sources by leveraging advanced mining tools manipulating and extracting context information out of selected metadata records.

Currently Data Searchery 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). Another planned extension deals with the possibility to plug-in additional data sources and extraction filter via simple property files instead of writing Java code. Data Searchery will be released to the OpenAIREplus community as an experimental real-time mining tool for end-users, as an addition to the pre-processing based inference services already implemented in the platform. The tool will automatically integrate repositories aggregated by the OpenAIRE infrastructure as DS data sources, hence make it possible for end-users to cross-search OpenAIRE repositories to discover relationships.

References

1. Bourne, P.E., Clark, T.W., Dale, R., de Waard, A., Herman, I., Hovy, E.H., Shotton, D.: Improving The Future of Research Communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331). Dagstuhl Manifestos **1**(1) (2012) 41–60
2. Hogenaar, A.: What is an enhanced publication? http://www.openaire.eu/en/component/content/article/76-highlights/344-a-short-introduction-to-enhanced-publications

3. Gray, J.: A transformed scientific method. In: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research (2009)
4. Reilly, S., Schallier, W., Schrimpf, S., Smit, E., Wilkinson, M.: Report on integration of data and publications. ODE Opportunities for Data Exchange
5. 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)
6. Chavan, V., Penev, L.: The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics **12**(S-15) (2011) S2
7. Hoogerwerf, M., Lösch, M., Schirrwagen, J., Callaghan, S., Manghi, P., Iatropoulou, K., Keramida, D., Rettberg, N.: Linking data and publications: Towards a cross-disciplinary approach. International Journal of Digital Curation **8**(1) (2013) 244–254
8. Castelli, D., Manghi, P., Thanos, C.: A vision towards scientific communication infrastructures - on bridging the realms of research digital libraries and scientific data centers. Journal on Digital Libraries (2013)
9. 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)
10. Manghi, P., Manola, N., Horstmann, W., Peters, D.: An infrastructure for managing ec funded research output - the openaire project. The Grey Journal (TGJ) : An International Journal on Grey Literature **6**(1) (2010) 31–40
11. Attwood, T.K., Kell, D.B., McDermott, P., Marsh, J., Pettifer, S.R., Thorne, D.: Utopia documents: linking scholarly literature with research data. Bioinformatics **26**(18) (2010) i568–i574
12. Feijen, M., Horstmann, W., Manghi, P., Robinson, M., Russell, R.: DRIVER: Building the Network for Accessing Digital Repositories across Europe. Ariadne **2007**(53) (October 2007)
13. Manghi, P., Mikulicic, M., Candela, L., Castelli, D., Pagano, P.: Realizing and maintaining aggregative digital library systems: D-net software toolkit and oaister system. D-Lib Magazine **16**(3/4) (2010)
14. Berners-Lee, T.: Linked data. http://www.w3.org/DesignIssues/LinkedData.html
15. Wölger, S., Siorpaes, K., Bürger, T., Simperl, E., Thaler, S., Hofer, C.: A survey on data interlinking methods. Technical report, Semantic Technology Institute (STI), University of Insbruck (March 2011)
16. Pipina, Karakos, A.S.: MusicPedia: Retrieving and Merging- Interlinking Music Metadata. International Journal of Computing **3**(8) (August 2011)