Context-driven Discoverability of Research Data

Miriam Baglioni$^{1[0000-0002-2273-9004]}$, Paolo Manghi$^{1[0000-0001-7291-3210]}$, Andrea Mannocci$^{1[0000-0002-5193-7851]}$

CNR-ISTI, Pisa, Italy
name.surname@isti.cnr.it

Abstract. Research data sharing has been proved to be key for accelerating scientific progress and fostering interdisciplinary research; hence, the ability to search, discover and reuse data items is nowadays vital in doing science. However, research data discovery is yet an open challenge. In many cases, descriptive metadata exhibit poor quality, and the ability to automatically enrich metadata with semantic information is limited by the data files format, which is typically not textual and hard to mine. More generally, however, researchers would like to find data used across different research experiments or even disciplines. Such needs are not met by traditional metadata description schemata, which are designed to freeze research data features at deposition time.

In this paper, we propose a methodology that enables "context-driven discovery" for research data thanks to their proven usage across research activities that might differ from the original one, potentially across diverse disciplines. The methodology exploits the collection of publication–dataset and dataset–dataset links provided by OpenAIRE Scholexplorer data citation index so to propagate articles metadata into related research datasets by leveraging semantic relatedness. Such "context propagation" process enables the construction of "context-enriched" metadata of datasets, which enables "context-driven" discoverability of research data. To this end, we provide a real-case evaluation of this technique applied to Scholexplorer. Due to the broad coverage of Scholexplorer, the evaluation documents the effectiveness of this technique at improving data discovery on a variety of research data repositories and databases.

1 Introduction

Over the last few years, research data have gained unprecedented importance and are now considered as central as traditional publications. Being able to search, find, access, and reuse such research products helps to accelerate scientific progress [3, 11], and cross-pollinate research by potentially fostering multidisciplinarity [10]. However, despite the extensive literature in the field of metadata-driven discovery technologies for scholarly communication, research data discovery still remains an open field of research. We can attribute these nonachievements to two main factors related to the yet immature positioning of research data in science.

Firstly, in many circumstances (e.g. "long-tail of data" scenarios), the absence of community practices, mandates, and incentives makes metadata description of research data unsatisfactory. Data is often perceived as supplementary material of an article and obtaining a persistent identifier (e.g. DOI) is the ultimate (and primary) aim of its deposition in a repository. To this end, research data metadata often do not undergo a curation and validation process as it occurs for libraries or publishers when research articles are submitted. Although the challenges hindering research data discovery seem in many ways similar to the ones arising for research articles and, more broadly, literature, the same solutions can hardly be applied. For example, the non-textual nature of data makes particularly hard the application of automated metadata enrichment techniques commonly in place when dealing with publications, such as natural language processing (NLP), full-text mining and topic extraction.

Secondly, research data discoverability is driven by user requirements that cannot be intrinsically satisfied by traditional metadata schemata/formats. While the discovery of research papers is motivated by the need of a researcher to find and read about the results of other scientists, the discovery of research data is driven by the need of finding data that can be reused to perform different analyses, in the same or even in different disciplines. Hence, even when research data are accurately deposited, and metadata is validated by data curators (e.g. thematic databases, repositories, archives), metadata structures cannot capture the variety of research applications the data may serve (or have subsequently served), and therefore fail in addressing such key discovery requirements. The semantic limits of metadata formats and, more broadly, the limits of the research data life-cycle, which disregards metadata enrichment based on further reuse, can be accounted as one of the main issues jeopardising data reuse practices and, ultimately, the enactment of open science.

In this work, as a solution to the problems above, we introduce the notion of *context-driven discoverability* of research data. The underlying intuition is that research data citation indexes, which populate an up-to-date graph of semantic relationships between research data and publications objects, can be exploited to propagate "research context", represented as a set of metadata properties of an object, to another related object. For example, the "abstract" and the "keywords" of an article metadata can be propagated and attached to the metadata description of research data being linked to the article via a relationship of type "cites". As a result of this process, the target research data, generated as an outcome of a given research activity and reused later to serve a different one, is also described by metadata that can leverage discovery by at least two distinct "research contexts".

To prove the effectiveness of context-driven discoverability, we *i)* present a *context propagation* technique for automated augmentation of bibliographic metadata of research data based on the semantic correlation between publications and data, and *ii)* perform an experimental study and validation of this technique using the OpenAIRE Scholexplorer's research data citation index.[1]

[1] Scholexplorer, `https://scholexplorer.openaire.eu`

Table 1: Scholexplorer entities and relationships.

Measure	Quantity
# of publications	21,288,342
# of datasets	51,946,754
# of relations publication-dataset	159,796,162
# of relations dataset-dataset (no loops)	141,403,762

Scholexplorer [4] aggregates and redistributes, free of charge, over 270 million bidirectional Scholix [5] links among research literature and datasets, and thus constitutes a fertile ground for our experimentation. Our experiment applies context propagation to the Scholexplorer citation graph showing how the resulting index can complete research data metadata and enable cross-context discovery of research data, across different research applications and across disciplines.

The remainder of the paper is structured as follows. Section 2 reviews Scholexplorer as primary data source for our experimentation, Section 3 describes our methodology to solve the problem, while Section 4 points out implementation details. Then, in Section 5 we evaluate our approach and discuss the results obtained, while Section 6 briefly reviews related work. Finally, in Section 7, we conclude and indicate possible extensions of our approach.

2 Data and resources

Having an up-to-date research data (also "dataset" in the following) citation index at disposal is a key enabling factor for this research. For our experiments we have relied on Scholexplorer, the OpenAIRE[2] service that provides over 270 million bidirectional Scholix [5] links among over 21 million research literature objects and 51 million datasets from 13,000 publishers, 10 data centres, CrossRef[3], Datacite[4], EMBL-EBI[5], and OpenAIRE. The whole collection is available, free of charge, via periodic dumps [7] and via API[6]. Table 1 shows the number of articles, datasets, and relationships in the dump used to perform our experiment.

The concept of *context* is flexible and may potentially include any relevant metadata field pertaining to publication entities, such as abstract, title, topics, keywords. The optimal setup might vary from dataset to dataset; indeed, a fine-tuning of the context to propagate can largely affect discoverability. In our experiment, we opted to propagate publication abstracts as they occur more frequently than topics and keywords, and therefore are a richer feed for full-text search. Besides, since relevant terms present in the title are generally present in the abstract too, we ruled out titles propagation.

[2] OpenAIRE, https://www.openaire.eu

[3] Crossref, https://www.crossref.org

[4] DataCite, https://www.datacite.org

[5] EMBL-EBI, https://www.ebi.ac.uk

[6] Scholexplorer API, https://scholexplorer.openaire.eu/#/api

Table 2: Potential impact of propagating abstract as context.

Measure	**Quantity**
# publications with abstracts	9,346,875
# datasets with abstracts	7,847,271
# rels between pubs with abst and dats	151,224,353
# rels between pubs with abst and dats with abst	5,288,025
# rels between pubs with abst and dats without abst	145,936,328

Table 3: Analysis of Scholexplorer subset of providers providing datasets. For each provider, the number of datasets is shown together with the relative percentage of datasets with abstract.

Provider	**Datasets**	**(% w/ abs)**
3TU.DC	164	(96.95%)
ANDS	29	(00.00%)
CCDC	716,009	(100.00%)
DataCite	8,470,681	(82.67%)
ENA	1,349,123	(42.36%)
ICPSR	6,823	(73.18%)
IEDA	488	(90.98%)
Pangaea	309,904	(38.53%)
RCSB	98,200	(00.00%)

To give a flavour of the impact of this choice, Table 2 reports on the total number of publications and datasets with abstracts and the number of relations from publications with abstract to datasets with or without an abstract. As can be noted, there is a significant number of relations (145,936,328) from publications with an abstract to datasets that could potentially benefit from context propagation. Table 3 completes this picture by reporting the number of datasets aggregated by Scholexplorer from each provider. It also highlights the percentage of datasets with a provided abstract, thus giving an indication on how “complete” are the potential targets of context propagation. Please notice, that a dataset (or a publication) in Scholexplorer can be potentially collected from several providers, hence, in this case, it would be counted multiple times.

The propagation process is driven by the semantics of the relationships between publication and data, and between dataset and dataset. Scholexplorer includes relationships whose semantics cannot be used for propagation, such as “hasMetadata”, which is not relevant to the research context; Table 4 provides a breakdown of the selected semantic relationships. Finally, given the selected subset of relationships, Table 5 reports the number of publications, datasets, and relationships (with and without abstracts) that are consequently involved in the propagation process.

3 Methodology

In this section, we introduce the terminology used in the paper and describe the chosen propagation strategy based on semantics. We define as:

Table 4: Breakdown of Scholexplorer selected semantics for context propagation.

	Semantics	**Quantity**
pubs–data	reviews	1,785
	references	1,949,635
	documents	258,513
	cites	169,397
	issourceof	30,052
	issupplementedby	1,238,320
	isderivedfrom	267

	Semantics	**Quantity**
data–data	isreferencedby	67,526,737
	isvariantformof	20,115
	references	67,526,737
	isdocumentedby	5,982
	continues	139,374
	documents	5,982
	haspart	1,178,496
	iscitedby	19,529
	issupplementedby	308,884
	isnewversionof	384,570
	cites	19,529
	issupplementto	308,884
	ispartof	1,178,496
	iscontinuedby	139,374

Table 5: Analysis of Scholexplorer subgraph according to the selected semantics.

Measure	**Quantity**
# of publications	1,065,121
# of datasets	4,886,298
# of relations (publication-dataset)	3,647,969
# of relations (dataset-dataset, no loops)	138,762,689
# publications with abstracts	574,209
# datasets with abstracts	3,392,081
# rels between pubs with abst and dats with abst	640,864
# rels between pubs with abst and dats without abst	1,788,183

Definition 1 (Context-driven discoverability) *The ability to discover a dataset based on information present in descriptive metadata of publications related to it, either directly (i.e. a publication refers this dataset) or indirectly (i.e. a publication refers a dataset that, in turn, refers this dataset, e.g. an earlier version of the same).*

Defined as such, context-driven discoverability essentially subsumes three possible scenarios of interest: *latent*, *reuse*, and *multidisciplinary* discoverability.

Definition 2 (Latent discoverability) *The ability to discover a dataset with incomplete metadata thanks to context propagated from another related object.*

Definition 3 (Reuse discoverability) *The ability to discover a dataset used for a research activity different from the one it has been created by, within the scope of the same disciplinary domain.*

Definition 4 (Multidisciplinary discoverability) *The ability to discover a dataset used for a research activity different from the one it has been created by, within the scope of a different disciplinary domain.*

All three scenarios covered by context-driven discoverability can be enabled by context propagation, which is defined as follows:

Definition 5 (Context propagation) *The process enabling context-driven discoverability. All the relevant semantic relations are followed in order to propagate context from publications so to form richer research data metadata records, which in turn propagate to other related research datasets. The process is limited by a threshold, defined by a termination function.*

The proposed methodology for context propagation relies on the fact that scholarly knowledge and research products (i.e. publications, research data, etc.) and their underlying relations can be represented as a graph. A *graph* is an ordered pair $G = (V, E)$ of nodes V and edges E. A *node* in the graph represents the abstraction of an entity in the modelled domain – in our case, a kind of research product (i.e. publications or research data) – while an *edge* represents a relationship between two nodes (e.g. a publication *reusing* a dataset). Nodes and edges can have labels that characterise them with attributes and specify their semantics. A source node u is said to be connected to a destination node v, indicated as $u \prec v$, when it exists an edge or an ordered set of edges (i.e. a *path*) connecting them.

The context propagation method here described relies on the existence of a path connecting two nodes, and on the chain of semantics connecting them, which reveals the reason for two nodes to be connected. For example, a publication could be connected to a dataset *directly* because the dataset *supplements* the publication (i.e. via an edge), or *indirectly* (i.e. through a path), e.g. because a newer version of a dataset *supersedes* the version originally *cited by* a publication (i.e. a path of length 2 exists from the publication to the newer dataset). The fact that two nodes are connected via a path allows us to propagate the context of a publication to relevant datasets. As already mentioned, the contextual information we chose to propagate to test our approach is the abstract.

The effect of context propagation depends on the "quality" of the path propagating the information from one node to another, which may depend on the semantics of the edges in the path or the length of the path. For this reason, our process associates a measure of *trust* to the propagated context that reflects the level of direct or indirect relatedness of the two nodes: the one propagating context and the one receiving it. Trust is key as it allows to filter out propagations with lower quality (i.e. a *cutoff threshold*), chose the most suitable propagation among many, or even set a termination function for the propagation process. Trust can be computed according to two strategies:

- Path-length driven: trust is inversely proportional to the length n of the path connecting two nodes, i.e. the shorter the path, the higher the quality. A trust function could be $1/n$. This case is trivial, and it is not an object of study in this paper.
- Semantic-driven: trust is mapped into a numerical weight characterising the edges of a path. The combination of such weights defines the trust of the relation between source and destination nodes. In this case, the trust can be

a number in the range $[0, 1]$ where 0 means no relatedness, and 1 means the maximum relatedness.

When the semantics of the relation is used to weight the edge connecting two nodes, the graph becomes a multi-graph, i.e. there could be multiple edges connecting two adjacent nodes. As edge semantics is a measure of the relatedness between two nodes, the higher the weight, the stronger the relation. Hence, the propagation strategy has to prefer paths that maximise the total weight. Given these premises, we define the propagation function as follows:

Definition 6 (Propagation function) *Given $G = (V, E)$ a multi-graph whose nodes belongs to two sets P (publications) and D (datasets), given $p \in P$ and $d \in D$ so that $p \prec d$, let w_{pd} be the maximum cumulative weight among all possible paths connecting the generic p at the generic d, and let $f_P(d) = PS_d = \{(p_i, w_{p_i d}) | p_i \in P \wedge p_i \prec d\}$ be the propagation function, which associates to d its propagation set (PS), where the generic weight $w_{p_i d}$ is such that:*

$$w_{p_i d} = \begin{cases} w_{d' d} * w_{p_i d'}, & (p_i, w_{p_i d}) \notin PS_d, \\ & (p_i, w_{p_i d'}) \in PS_{d'}, \\ & (d', d) \in E \\ \\ max(w_{p_i d}, (w_{d' d} * w_{p_i d'})), & (p_i, w_{p_i d}) \in PS_d, \\ & (p_i, w_{p_i d'}) \in PS_{d'}, \\ & (d', d) \in E \\ \\ w_{p_i d}, & (p_i, d) \in E \end{cases}$$

The propagation function depends on the product of the semantic relatedness weights in the path, and always prefers the edges with the highest weight among those at its disposal in the chosen path. Among all the computed paths connecting a couple of nodes, it chooses the path maximising the overall weight independently from its length. In this way, a low semantics relatedness along a path plays an important role as a discount factor and helps to filter unsatisfactory propagations out. At the same time, it does not penalise long paths with strong semantic relatedness.

Figure 1 shows an example of the propagation process over a sample graph. On the left-hand side, Figure 1a shows the graph in its starting condition before propagation takes place: blue nodes refer to publications, red ones to datasets, and the edge associated to the semantics with maximum weight between each couple of nodes is shown. For simplicity, we assume a semantic relation and its inverse have the same weight. We also fix the trust cutoff threshold to 0.3. In each iteration, all the nodes with available context for propagation are considered and try to affect all their neighbours. In the first step, only publications have a context at disposal for propagation, so all the edges connecting them to datasets are considered (represented as dashed in the figure). This is shown in Figure 1b: both D_1 and D_3 receive context respectively by P_1 and P_2. Each propagated context has the same weight as the edge involved since it is a direct connection. In any step other than the first one, the context is propagated among datasets. Each dataset having received a previously propagated context

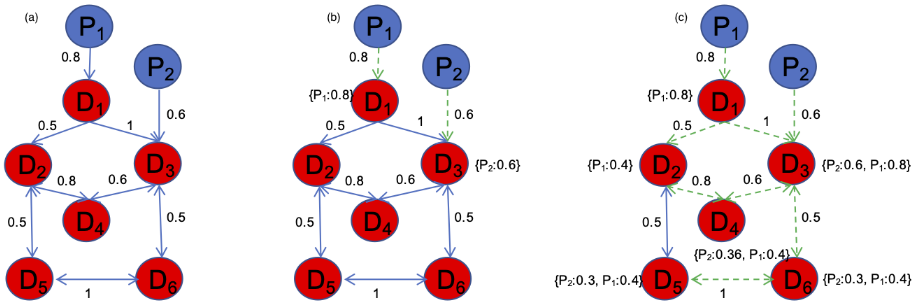

Fig. 1: Context Propagation example.

tries to pass it along to all its neighbouring datasets. However, this time the weight of the association is not equal to the weight of the edge connecting the dataset with its neighbours. In fact, the context has been "inherited" from a publication, and thus the indirect connection has to be taken into account. Each time a context is further propagated between two datasets d' and d, its weight is computed by multiplying the weight for the context seen at d' and the weight of the edge connecting d' and d. A context is propagated to a dataset only if it does not already belong to the dataset's PS. In case the PS already contains information about the publication whose context is being passed, its weight is computed as the maximum among the weights computed on the paths that have reached the node so far. Figure 1c shows the graph after the propagation process has terminated. D_1 receives context only from P_1, D_3 receives context from P_2 directly, and from P_1 through D_1. The weight for the context of P_1 is the same for both the dataset, since the edge that binds them weights 1. D_2 receives context of P_1 through D_1 and the strength of the correlation is multiplied by 0.5 (i.e. the edge weight). It does not receive context from P_2 because it could propagate only through D_4 or D_5, but the strength of the relation would be below the cutoff threshold in both cases, and thus they are discarded. D_4 receives propagation information from both P_1 and P_2, and both of them through D_3: the propagation weight of P_1 through D_1 would have been 0.32, which is less than the 0.4 got from D_3. D_6 also receives propagation information for both P_1 and P_2 through D_3, and the correlation strength is multiplies by 0.5 for both the publications. D_6 receives propagation information from P_1 and P_2 through D_5 with the same weight of D_5.

4 Implementation

As Scholexplorer dump occupies over 40 GB compressed on disk, it was unfeasible to treat the problem with an in-memory approach. We opted for the utilisation of our Hadoop[7] cluster and implemented the propagation algorithm as a sequence of Spark jobs in PySpark. The code is publicly accessible here[8].

Running context propagation takes about 6 hours on our cluster with 20 virtual machines (VMs) for Apache HDFS DataNodes and Spark workers, each VM with 16 cores, 32 GB of RAM, and 250 GB of space on disk; plus 3 dedicated virtual machines for HDFS Name Nodes, each one with 8 cores, 16 GB of RAM, and 40 GB of space on disk. Please notice that our termination function on the Scholexplorer graph makes the process terminate after three steps of propagation, that is one direct propagation from publications to datasets, and just two steps of propagation between datasets. We do believe this is acceptable for the sake of computational feasibility as the number of nodes reachable by context propagation after three steps covers about 97% of the nodes reachable via paths originating from publications with context (evaluated through an iterative graph exploration converging at 2,266,269 nodes).

In order to make the evaluation of the proposed approach easier, we provide two full-text indexes on Elasticsearch[9]. A first index (*propagation-before*) contains metadata records from Scholexplorer before context propagation has run, while a second one (*propagation-after*) provides the same records after context propagation is performed. The former contains metadata of publications and datasets consisting mainly of *identifier*, *pid*, *type* (i.e. publication or dataset), *title*, *abstract*. The latter contains the same metadata descriptions plus one more field (*propagated abstracts*) for datasets in order to amass the abstracts coming from publications via context propagation. In order to evaluate the results, the user can play with a simple search interface[10] and explore the saved queries as examples (refer to Section 5.2), or query the indexes from scratch.

5 Evaluation

In this section, we present the results obtained by applying the methodology proposed in Section 3 and characterise them both from a quantitative and qualitative standpoint.

5.1 Quantitative analysis

As a mean of comparison, Table 6 reports on the number of datasets and relative percentage of datasets with abstract for each provider involved in the analysis. Table 7 instead reports the results obtained by the application of the context propagation. For each provider, the table shows the number of datasets affected by context propagation both directly from a paper ("Publication–Data" column) and indirectly from another dataset ("Data–Data" column). For each provider

[7] Hadoop, `https://hadoop.apache.org`

[8] Code repository, `https://code-repo.d4science.org/miriam.baglioni/context-propagation`

[9] Elasticsearch, `https://www.elastic.co/elasticsearch`

[10] Evaluation interface: `https://propagation-demo.infrascience.isti.cnr.it`

Table 6: Analysis of Scholexplorer subset of providers providing datasets in the subgraph selected according to the valid semantics. For each provider, the number of datasets is shown together with the relative percentage of datasets with abstract.

Provider	**Datasets**	**(% w/ abs)**
3TU.DC	62	(93.55%)
ANDS	2	(00.00%)
CCDC	713,350	(100.00%)
DataCite	3,796,690	(88.52%)
ENA	339,868	(00.00%)
ICPSR	6,823	(73.18%)
IEDA	443	(99.32%)
Pangaea	150,759	(45.88%)
RCSB	70,557	(00.00%)

Table 7: Quantitative evaluation of context propagation. For each provider, the number of datasets touched by propagation is reported together with an estimation of latent and reuse discoverability.

Provider	**Publication–Data**				**Data–Data**			
	Propagated contexts	**(% tot)**	**Latent**	**Reuse**	**Propagated context**	**(% tot)**	**Latent**	**Reuse**
3TU.DC	27	(43.55%)	0	15	12	(19.35%)	0	8
ANDS	1	(50.00%)	1	0	–		–	–
CCDC	130,317	(18.27%)	0	333	546	(0.08%)	0	225
DataCite	405,088	(10.67%)	4,921	28,619	849,260	(22.37%)	24,859	656,862
ENA	337,814	(99.40%)	337,814	60,888	–		–	–
ICPSR	3,691	(54.10%)	743	3,303	130	(1.91%)	4	78
IEDA	41	(9.26%)	1	7	16	(3.61%)	0	6
Pangea	2,951	(1.96%)	200	600	35,770	(23.73%)	12,571	10,200
RCSB	70,398	(99.77%)	70,398	46,133	–		–	–

it shows *i)* the total number of datasets receiving an abstract ("Propagated contexts") and the percentage relative to the total reported in Table 6, *ii)* an estimation of latent discoverability ("Latent" column), and iii) an estimation of reuse discoverability ("Reuse" column). Latent discoverability is evaluated by counting the number of datasets without own abstract that have been targeted by context propagation, while reuse is evaluated by counting the number of datasets receiving at least two propagated contexts. Please note that the reuse estimation computed as such incurs in an underestimation of the potential reuse. In fact, as an example, it does not account for datasets whose only semantic relation available is the one connecting the dataset to the publication reusing it (i.e. it should be accounted for reuse, but in this case only one context is propagated, and thus it is not counted). From the reported results, we can notice that the margin of improvement varies largely across providers: from a cumulative (i.e. for all propagation steps) 12.87% for IEDA to almost the totality for ENA and RCSB (99.40% and 99.77% respectively). We believe that focusing on absolute numbers does not deliver the right key to interpret the results as a seemingly marginal improvement can still be significant in terms of discoverability for users.

Moreover, providing a quantitative estimation of multidisciplinary discoverability is extremely hard as there are no objective tests to identify such cases. For most of them, only a domain expert has the in-depth knowledge to judge whether it is truly multidisciplinary or not; for this reason, we study this aspect from a qualitative standpoint only.

Finally, it is worth mentioning, that a few providers do not participate in data–data propagation as their datasets are not related with the selected semantics for our experimentation. For example, nucleotides provided by ENA do not have relations among them, but only towards publications mentioning them, thus they do not participate in data–data propagation.

5.2 Qualitative analysis

In this section, we present a collection of a few chosen examples that, to the best of our knowledge, better describe from a qualitative standpoint the results achieved through our approach and advocate for its application. The reader can find them via the evaluation interface provided.

Example 1 (Latent discoverability) *The query term "SHC014" is the name of a coronavirus spike protein that has recently resonated in the media worldwide; resulted from a 2015 lab experiment, it has been wrongly associated to the current SARS-CoV-2 outbreak. The query term in the original index matches only the publication relative to the original experiment, while after the propagation the dataset "Structure of SARS coronavirus spike receptor-binding domain complexed with its receptor" emerges, despite being originally deprived of any further metadata, but the title. Moreover, as can be seen, several other relevant publication abstracts are included as dataset context, hence improving its discoverability dramatically.*

Example 2 (Reuse discoverability) *The YfdE gene from the bacteria E.Coli receives the abstracts of two publications thanks to context propagation: the first is a 2013 publication describing gene's function and x-ray crystal structure, while the second one is a 2018 paper referring the protein acetyl-CoA:oxalate CoA-transferase, which the gene synthesise.*

Example 3 (Multidisciplinary discoverability) *We managed to isolate the "PRIMAP-hist Socio-Eco dataset" which is used relevant to both assess anthropogenic land-use estimates and for the creation of a consistent historical time series of GDPs for 195 countries in the last 150 years. As anticipated, pinpointing true multidisciplinary examples is rather difficult without prior domain knowledge. However, we believe this example can still be a good candidate that shows how two disciplines within Earth System Science can benefit from context-driven discoverability.*

6 Related Work

The approach described in this paper and the problems it addresses share a few peculiarities with other research problems from other research applications.

A first similar application is Automated Query Expansion (AQE), whose major contributions across over 50 years of research are reviewed and summarised in [1,6]. In AQE, the terms composing the user query are expanded by adding a new set of features at query time by means of different techniques (e.g. stemming, dictionary and ontology-based augmentation, language modelling, query rewrite) in order to capture a broader set of potentially relevant documents (i.e. improve recall, generally, at the expense of decreasing precision). However, this is seldom effective in our case, as there is often little to be matched in research data descriptions. Indeed, research data metadata are often largely incomplete, and so, even if the user query is automatically-expanded consistently, the search seldom can retrieve further results potentially relevant for the user. To some extent, our approach can be still categorised as an augmentation task as in AQE. In fact, rather than augmenting the terms contained in the user query thanks to language models, we augment the metadata descriptions in research data by propagating information following their semantic relations towards relevant literature and other research data. Unlike AQE techniques, where the user might be puzzled when trying to understand why certain documents have been returned with high saliency despite being very different from the expressed query terms, our method can always provide the user with the information needed to explain why a given result has been returned as potentially relevant. In a similar way to AQE, an early work from Mannocci et al. addressed research data discoverability by providing a user interface enabling the composition of on-the-fly queries against research data archives starting from a literature record of interest [9].

A second similarity is shared with Label Propagation (LP) [14–16]. Within the research field of complex networks, LP is a specific task that aims at labelling a large quantity of unlabelled nodes across the network starting from the little knowledge present in a much smaller group of labelled nodes. Such labelling is in practice performed by propagating a finite set of labels across the network by means of nodes properties and their semantic relations (i.e. network topology). Such algorithms are originally devised to detect communities in networks, but nonetheless, they share to some extent common properties with the class of problems introduced in this work. A typical case study for such class of algorithms is the propagation of political affiliation in Fiend-of-a-friend (FOAF) networks (i.e. identify communities or clusters of right-wing and left-wing nodes). Like LP, our context of applications deals with nodes rich in information (i.e. labelled) and nodes poorer in information (i.e. unlabelled); however, in our case, the split among the two classes is far more balanced than the one noticeable in typical LP applications. Similarly, LP indeed tries to spread information (i.e. the labels) from one node to another; however, unlike in our task, the set of candidate labels is finite and known *a priori* (e.g. in the case of political orientation: “right-wing” or “left-wing”). In our application instead, the amount of information the algorithm can potentially propagate across the network is not known *a priori* and, in general, grows with the size of the network (i.e. one unique abstract for each publication joining the network). Indeed, any publication node could offer its own "label value" as a propagation candidate; however, we cannot talk about community detection in our case study as there is no real community to be discovered.

7 Conclusions and Future Work

In this paper, we described a sound methodology enabling context-driven discoverability for research data thanks to their proven usage across research activities that might differ from the original one, potentially across diverse disciplines. We showed how publication–dataset semantic relations can be leveraged in order to propagate research context (e.g. abstracts) from publication to dataset, and thus form richer metadata description. By providing a real-case evaluation on Scholexplorer, we showed how a large number of datasets across all Scholexplorer providers can benefit from the context propagated from related literature, and showcased a few selected representative examples.

The context propagation methodology here proposed can be improved and refined in several different directions. During our experiments, we observed that some semantics can be more conducive for a type of discoverability (i.e. latent, reuse, multidisciplinary) than for the others. For example, semantics as *isSupplementedBy*, *documents* or *reviews* between publication and dataset strongly suggest a potential case of latent discoverability within the scope of the same research application, while *cites* or *references* can indicate most probably a reuse. To this end, semantics could be tightly associated with the three different types of discoverability by providing a different weight for each one of them.

Moreover, in order to assess further the capabilities in multidisciplinary research, and isolate better candidates that are difficult to retrieve otherwise (especially without in-domain knowledge), we could leverage topics and keywords along with abstracts. This would enable us to match topics with known ontologies such as MeSH [8] for Life Sciences, PhySH [13] in Physics, CSO [12] for Computer Science, and therefore gain a better view on whether a dataset effectively lies on the border of two (or more) disciplines. More sophisticated NLP techniques, such as Latent Dirichlet Allocation [2], could be applied in order to let latent structure emerge from abstract plain-texts and characterise further the nodes alongside topics and keywords.

Furthermore, it is in our plans to study the feasibility of an extensive search-based user evaluation by providing access to the propagated index so to log user queries and interactions with the results (e.g. relevant, not relevant). Such knowledge can be used as ground truth in order to accurately assess the improvement achieved by context propagation by rerunning the same queries under the hood against the other index and measure the differences.

Finally, in order not to disperse the added value, propagated information could be fed back to content providers, so that it can be integrated into the original data catalogues so to deliver context-driven discoverability out-of-the-box right where it belongs and can be more effective.

Acknowledgements

This work was co-funded by the EU H2020 project OpenAIRE-Advance (Grant agreement ID: 777541).

References

1. Bhogal, J., MacFarlane, A., Smith, P.: A review of ontology based query expansion. Information processing & management **43**(4), 866–886 (2007)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
3. Borgman, C.L.: Big Data, Little Data, No Data. The MIT Press (2015). https://doi.org/10.7551/mitpress/9963.001.0001
4. Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M., Schindler, U.: The data-literature interlinking service: towards a common infrastructure for sharing data-article links. Program **51**(1), null (2017). https://doi.org/10.1108/PROG-06-2016-0048
5. Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M., Schindler, U.: The scholix framework for interoperability in data-literature information exchange. D-Lib Magazine **23**(1/2) (2017)
6. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. Acm Computing Surveys (CSUR) **44**(1), 1–50 (2012)
7. La Bruzzo, S., Manghi, P.: Openaire scholexplorer service: Scholix json dump (Nov 2019). https://doi.org/10.5281/zenodo.3541646
8. Lipscomb, C.E.: Medical subject headings (mesh). Bulletin of the Medical Library Association **88**(3), 265 (2000)
9. Mannocci, A., Manghi, P.: Preliminary analysis of data sources interlinking. In: International Conference on Theory and Practice of Digital Libraries. pp. 53–64. Springer (2013). https://doi.org/10.1007/978-3-319-08425-1_6
10. Pasquetto, I.V., Borgman, C.L., Wofford, M.F.: Uses and Reuses of Scientific Data: The Data Creators' Advantage. Harvard Data Science Review **1**(2) (nov 2019). https://doi.org/10.1162/99608f92.fc14bf2d
11. Pasquetto, I.V., Randles, B.M., Borgman, C.L.: On the reuse of scientific data. Data Science Journal **16**(Borgman 2015), 1–9 (2017). https://doi.org/10.5334/dsj-2017-008
12. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: a large-scale taxonomy of research areas. In: International Semantic Web Conference. pp. 187–205. Springer (2018)
13. Smith, A.: Physics subject headings (physh). ISKO Encyclopedia of Knowledge Organization (2019)
14. Wang, F., Zhang, C.: Label propagation through linear neighborhoods. IEEE Transactions on Knowledge and Data Engineering **20**(1), 55–67 (2007)
15. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in neural information processing systems. pp. 321–328 (2004)
16. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International conference on Machine learning (ICML-03). pp. 912–919 (2003)