Context-driven discoverability of research data

Miriam Baglioni, Paolo Manghi, **Andrea Mannocci**

Consiglio Nazionale
delle Ricerche

ISTITUTO DI SCIENZA E TECNOLOGIE
DELL'INFORMAZIONE "A. FAEDO"

infraScience
infrastructures4ScienceGroup

Outline

- Introduction and context
- Research question
- Data and resources
- Methodology
- Conclusion and future work

Introduction

Research data have **pivotal importance** in nowadays research

- Searchability
- Findability
- Accessibility
- Reusability

… so to foster

- Acceleration of scientific progress
- Reproducibility of science
- Cross-pollination of research and multidisciplinarity

Introduction

Yet, research data positioning is **rather immature** in scholarly communication and, more broadly, in science

- Absence of
  - common practices (e.g. at community level)
  - incentives for researchers
  - mandates/policies
- Often perceived as **ancillary material**
  - e.g. DOI minting is the ultimate goal
- In terms of curation, it cannot undergo the same mining approaches that would work with literature (i.e. fulltext)

Introduction

Moreover, research data discoverability is driven by user requirements that cannot be intrinsically satisfied by traditional metadata schemata/formats and procedures.

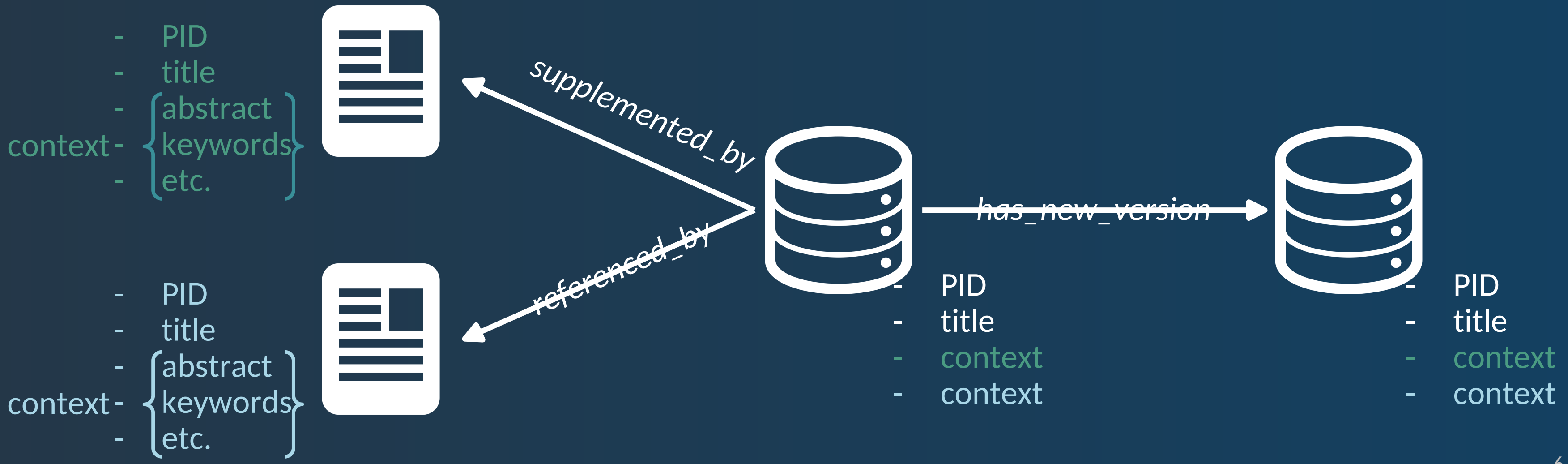
- Substantial difference from literature

Discover and read results by other researchers

Discover data that can be reused so to perform different analyses

The idea: context-driven discoverability

Intuition: leverage the network of semantic relations among literature and research data objects in data citation indices, so to move "research context" from literature to research data and propagate it further.



Data and resources

Benchmarking dataset: we opted for **OpenAIRE Scholexplorer** as the dataset of reference for this work, http://scholexplorer.openaire.eu

Preprocessing:

- prune isolated publication and datasets (no relation)
- select a number of semantic relations of interest for context-driven discoverability (based on DataCite schema, https://schema.datacite.org/meta/kernel-4.3/doc/DataCite-MetadataKernel_v4.3.pdf)

YES: *is_supplemented_by, is_new_version_of*

NO: *compiles, has_metadata*

Data and resources

Benchmarking dataset: we opted for **OpenAIRE Scholexplorer** as the dataset of reference for this work, http://scholexplorer.openaire.eu

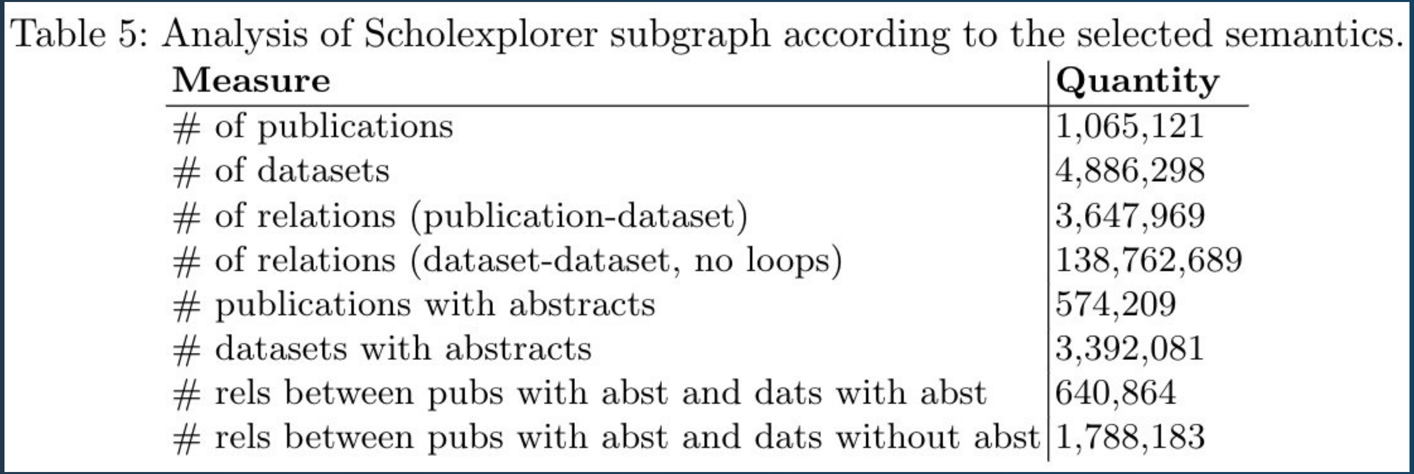


Table 5: Analysis of Scholexplorer subgraph according to the selected semantics.

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

Data and resources

Benchmarking dataset: we opted for **OpenAIRE Scholexplorer** as the dataset of reference for this work, http://scholexplorer.openaire.eu

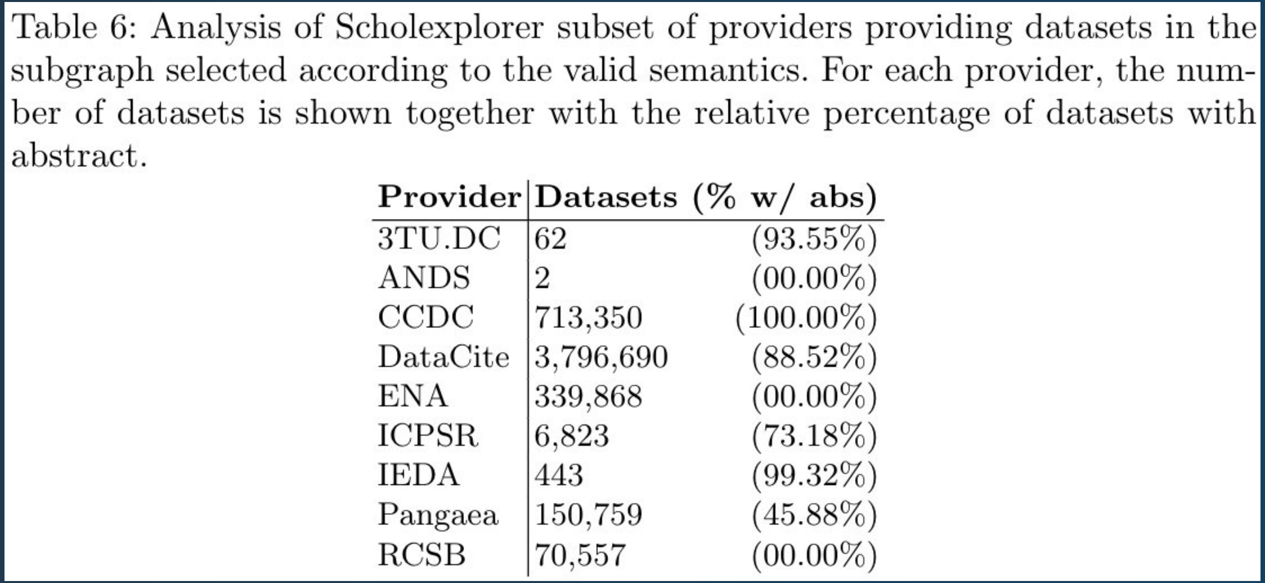


Table 6: Analysis of Scholexplorer subset of providers providing datasets in the subgraph selected according to the valid semantics. For each provider, the number of datasets is shown together with the relative percentage of datasets with abstract.

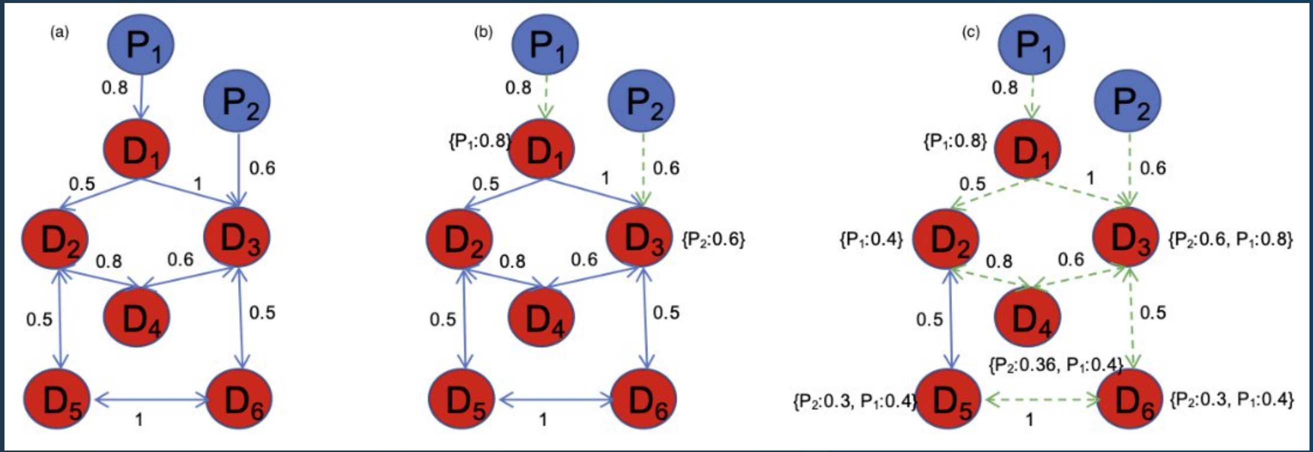
Provider	Datasets	(% w/ abs)
3TU.DC	62	(93.55%)
ANDS	2	(00.00%)
CCDC	713,350	(100.00%)
DataCite	3,796,690	(88.52%)
ENA	339,868	(00.00%)
ICPSR	6,823	(73.18%)
IEDA	443	(99.32%)
Pangaea	150,759	(45.88%)
RCSB	70,557	(00.00%)

Methodology

- **Context-driven discoverability** comes in three different "profiles"
  - **Latent**: ability to discover a dataset with incomplete metadata thanks to context propagated from another related object.
  - **Reuse**: 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
  - **Multidisciplinary**: 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
- **Context propagation**: process enabling context driven discoverability
- Arbitrary choice → **context ≔ abstract**

Methodology - Context propagation

- Context is propagated along paths from publication to data
  - In rounds
  - **Weights** are statically assigned to edges according to the semantics
  - **Cumulative weight** is computed along the path (i.e. products)
  - A **cutoff threshold** to discard low quality propagated contexts



Implementation

- **40 GB compressed** dataset → in-memory approach unfeasible
- Implemented as a **Spark** job in PySpark and run on our **cluster**
  - 20 virtual machines (VMs) for Apache HDFS DataNodes and Spark workers
    - 16 cores
    - 32 GB of RAM
    - 250 GB of space on disk
  - 3 dedicated virtual machines for HDFS Name Nodes
    - 8 cores
    - 16 GB of RAM
    - 40 GB of space on disk
- **6 hours completion time** for the job
  - <u>3 steps</u> of propagation (1x publication → data, and 2x data → data)

Quantitative evaluation

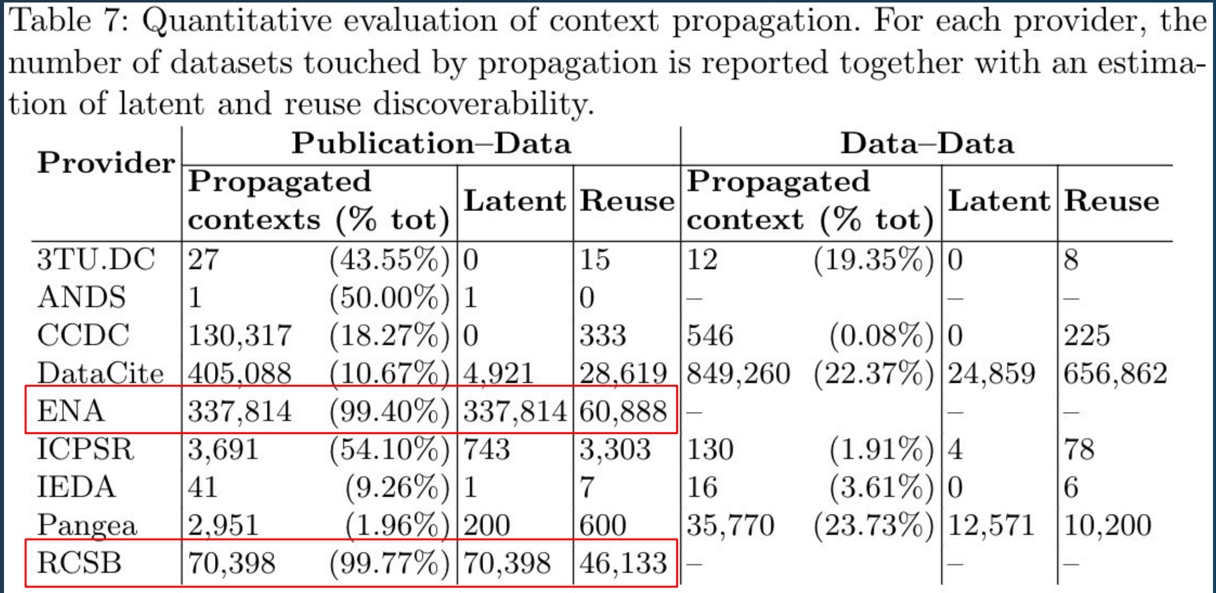


Table 7: Quantitative evaluation of context propagation. For each provider, the number of datasets touched by propagation is reported together with an estimation of latent and reuse discoverability.

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	–	–	–

Noteworthy: finding significant examples of multidisciplinary research is an arduous task without knowing a priori what to look at, or, in general, without in-depth domain knowledge.

Qualitative evaluation

- Offer an interface for inspection of records, before and after context propagation
- We isolated three records as example, one for each kind of context-driven discoverability (i.e. latent, reuse, multidisciplinary).

Live demo: https://propagation-demo.infrascience.isti.cnr.it

Conclusions and future works

- We suggested how research data discoverability can be improved via context-driven discoverability
- We presented a methodology enabling context propagation by leveraging the presence of semantic relations among data and literature objects so to propagate contextual information
- Future extensions
  - Study the feasibility of an user-based evaluation
  - Tie different semantics to the three different discoverability profiles
  - Leverage keywords and topics to support further the identification of potential multidisciplinary candidates
  - Align to existing topic ontologies (e.g. MeSH, PhySH, CSO, etc.)
  - Apply more sophisticated NLP techniques (e.g. LDA)

Consiglio Nazionale
delle Ricerche

ISTITUTO DI SCIENZA E TECNOLOGIE
DELL'INFORMAZIONE "A. FAEDO"

Thank you!

Any question?

Andrea Mannocci
ISTI-CNR
andrea.mannocci@isti.cnr.it