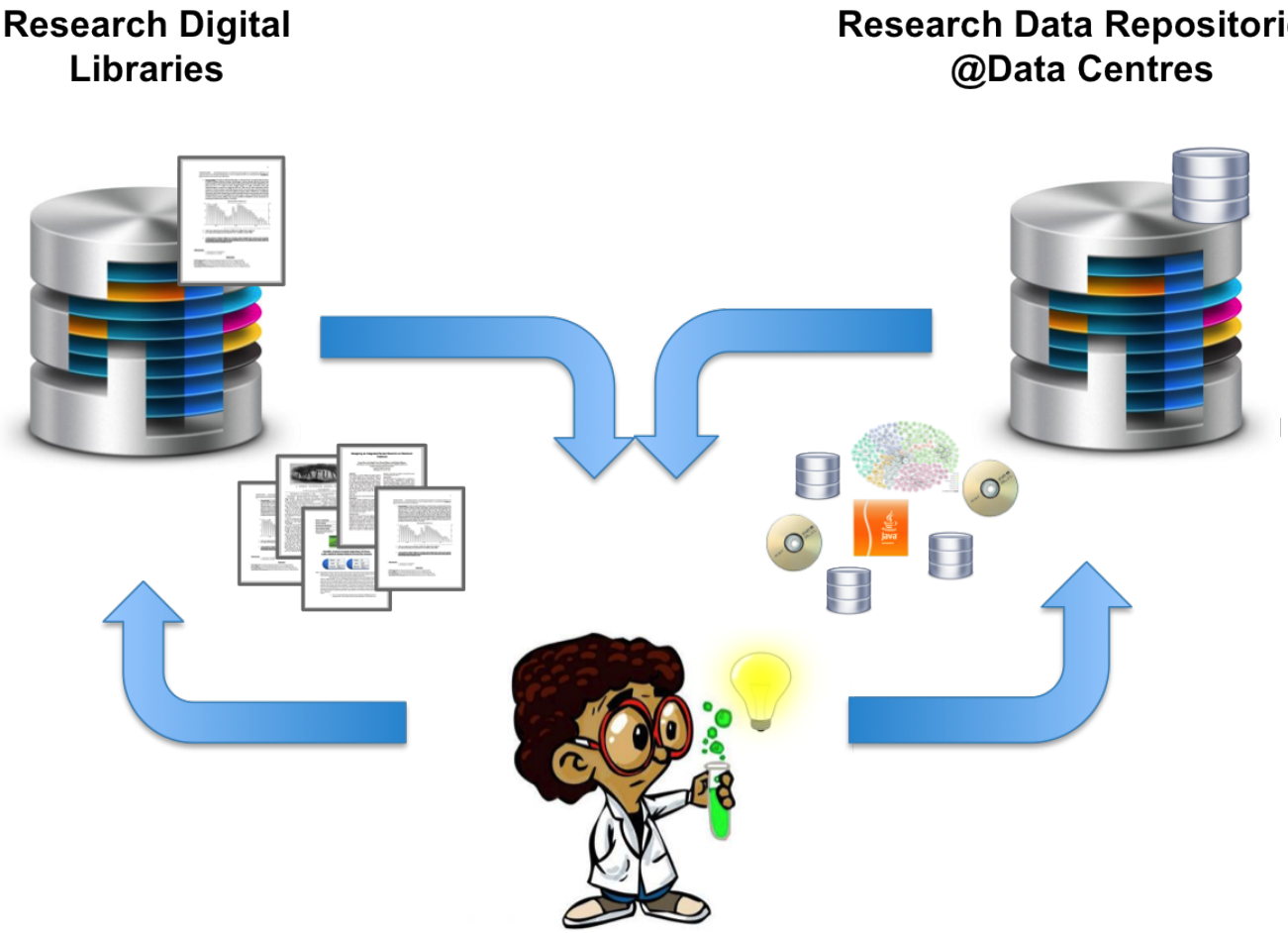
iCORDI
INTERNATIONAL COLLABORATION
ON RESEARCH DATA INFRASTRUCTURE

DATA SEARCHERY

Preliminary Analysis of Data Sources Interlinking

Andrea Mannocci and Paolo Manghi
ISTI-CNR

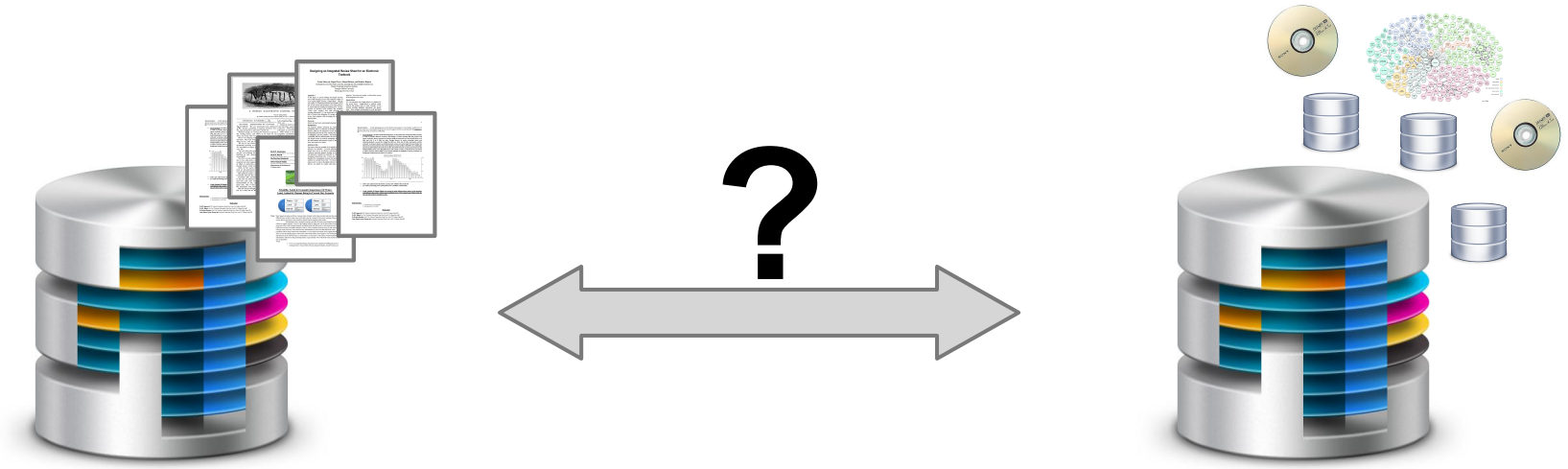
Modern eScience workflow



DATA SEARCHERY

Modern eScience workflow

Lack of tools for data-publication interlinking



Research Digital Libraries

Research Data Repositories

Benefits:

- Foster multidisciplinary research by looking at adherences among distinct disciplines
- Enable better review, understanding, reproduction and re-use of research activities

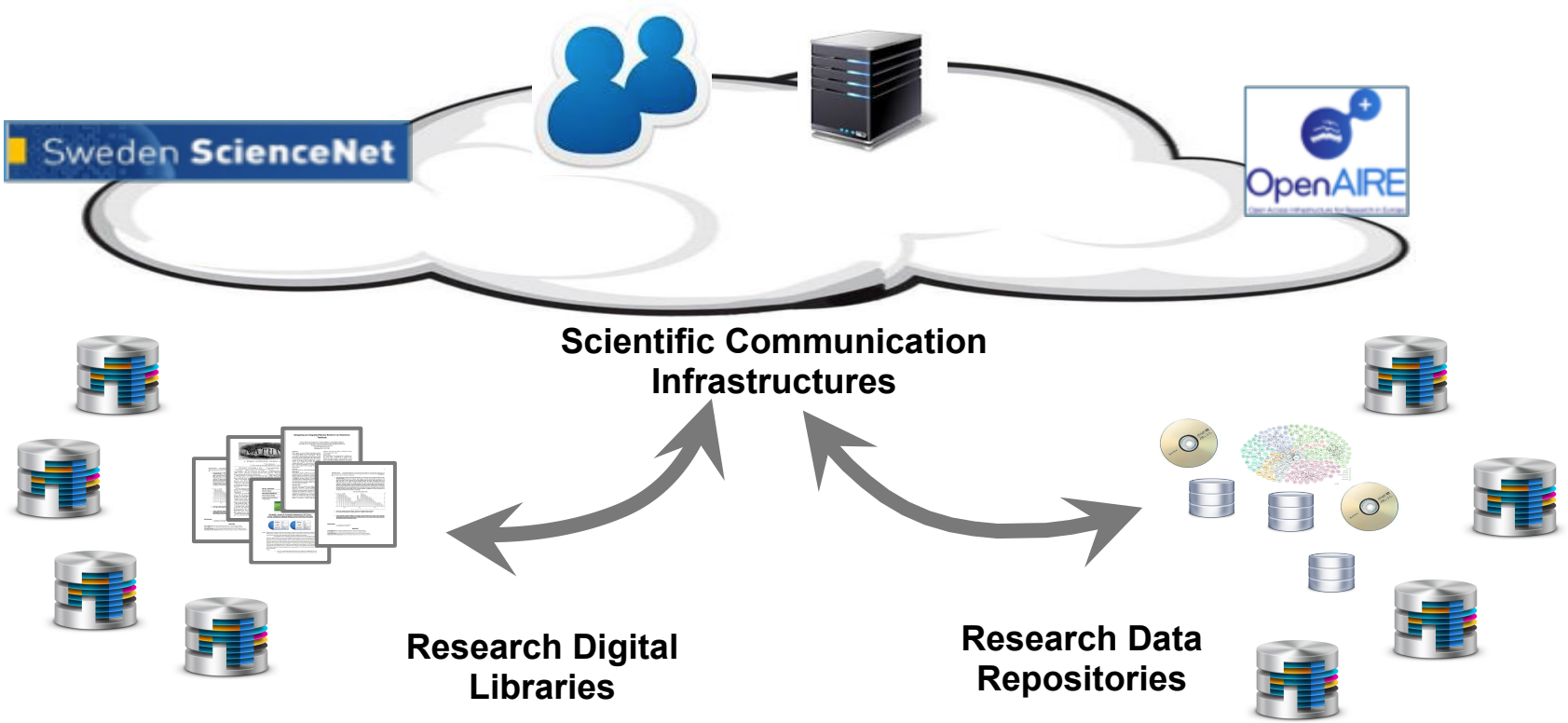
DATA SEARCHERY

Scientific Communication Infrastructures

Interlinking and contextualizing publications and data sets

Services and tools for

- **Aggregation of content** (e.g. harvesting, harmonization, inference, editing)
- **Provision** (e.g. web portals, standard APIs)



Scientific Communication Infrastructures

Drawbacks

- High costs for design and development
  - Ever changing requirements from case to case and over time
  - Long time-to-deployment
  - Critical maintenance procedures
- High costs of operation
  - Data curation
  - Data inference

The idea

DATA SEARCHERY

Design a tool...

- Light
- Flexible

...enabling users to **surf and** (**best-effort**) **relate on-the-fly** metadata present in two different web data sources.

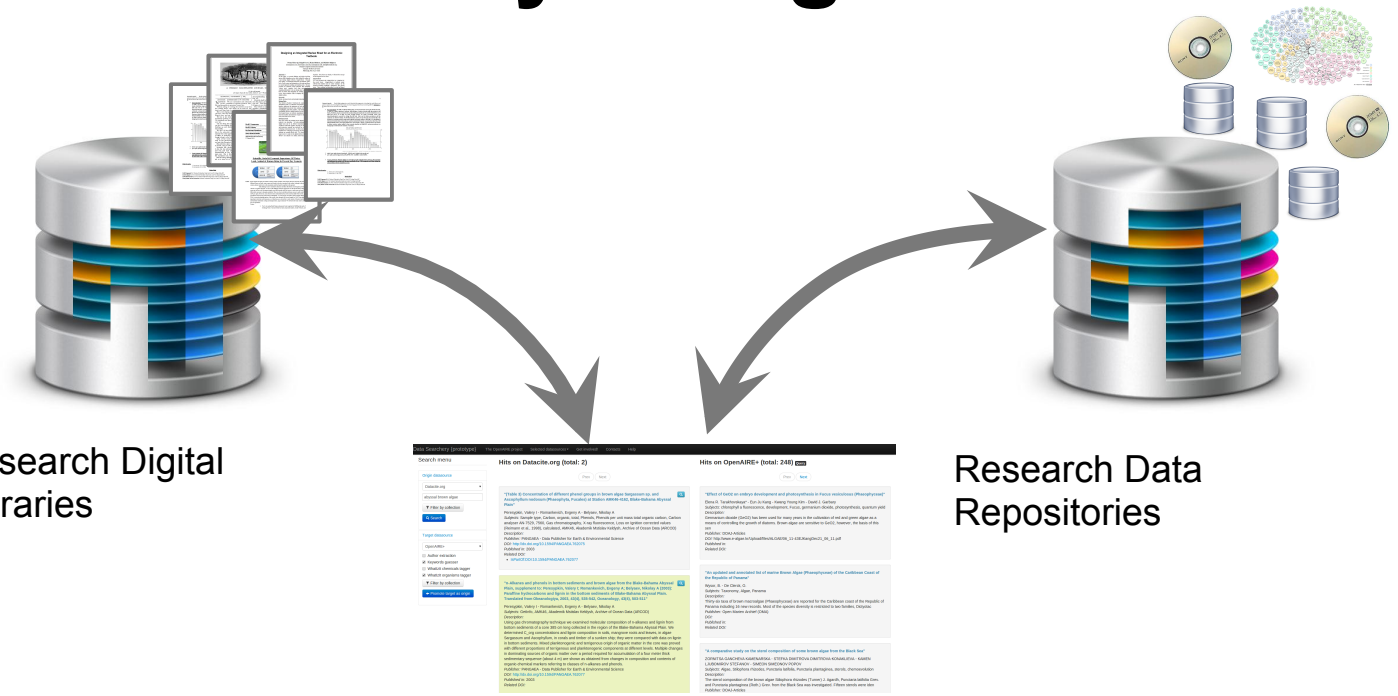
In such a way:

- Unneeded costs (of aggregation) during SCIs development can be cut
- Users can search for and play with metadata even if a SCI is not yet ready

Not only!

- It can be used as an alternative to SCI, whenever SCIs are not affordable
- It can be integrated to existing SCIs as an additional tool for mining

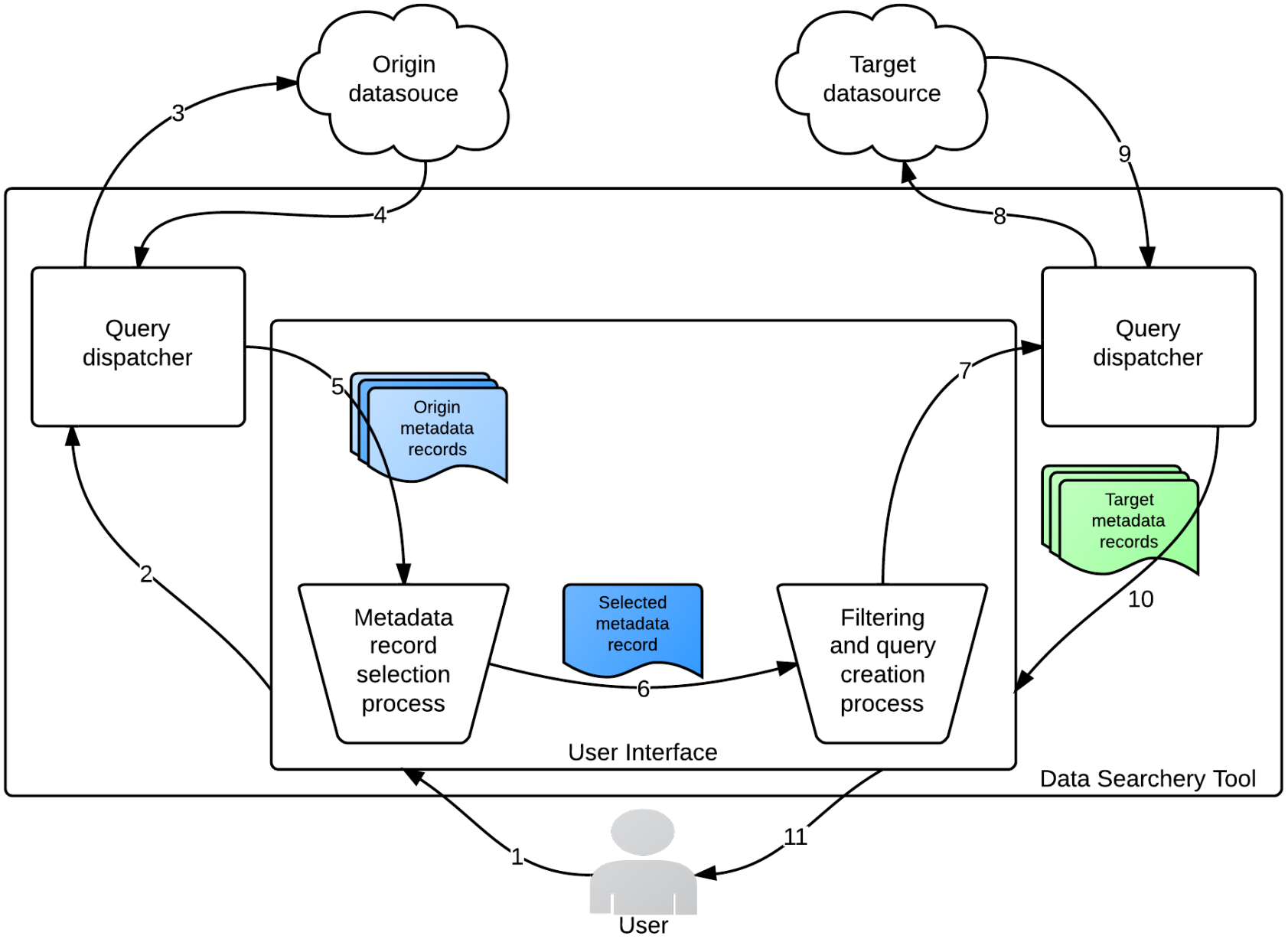
Data Searchery at a glance



- Data searchery **just runs real-time queries** on web data sources: no metadata harvesting, nor pre(post) processing takes place.
- Data Searchery **combines the textual query with information extracted from selected metadata** fields thanks to extraction filters.
- With Data Searchery an user can query two data sources and interlink their objects in **just one browser tab**.

DATA SEARCHERY

Data Searchery at a glance



Data Searchery

Main actors in play

Data Source

- Export of XML-formatted metadata
- Apache Sorl web search api
- Optionally organized into collections

Extraction Filter

- Keywords extraction from metadata fields
- Implementation can be
  - local
  - remote (demanded to external web services, e.g. whatizit, text tagger services, etc.)

Data Searchery

Extendibility considerations

Data Searchery can be easily customized by adding a few classes

- New data sources
- New extraction filters

Data Searchery

An example

1. Select an origin data source out of the ones implemented (say Datacite.org)
2. Search for some keyword (let's go for "calcification foraminifer")
3. Select a target data source (say OpenAIRE+) and check out "Author filter"
4. Choose a record and click on the magnifying glass
5. Check the right column for results!

Data Searchery

Testing results

- The tool in its current version helped us in finding and confirming some linked publications and datasets within the OpenAIREplus infrastructure.

- **Alas.. no epiphanies!**
  - Data Searchery works better if you somehow have some **prior understanding** on what's inside repositories.
  - Finding totally unexpected relationships given whatsoever queries and two random data sources is seldom.. (so far!)
- Furthermore, the **recall** of the approach is proportional to:
  - how **rich and accurate metadata** records are
  - how **good filters** have been implemented
  - how much **cohesion** there is between two data sources

Future work

DATA SEARCHERY

Enhancements

- More precise implementation of extraction filters
- Deliver to the user a fine-grained control over the generated query

Extensions

- Bulk analysis of correlation of data sources
  - Definition of sets of queries to analyse correlation
  - Identifying measures of “potential correlation”
- Implement new backends for query (e.g. ElasticSearch, JDBC, OpenSearch)
- Integration in OpenAIRE as an extension

Questions?

Feel free to contact us!!

Andrea Mannocci and Paolo Manghi

{andrea.mannocci, paolo.manghi}@isti.cnr.it

InfraScience Research Group

ISTI-CNR, Pisa, Italy

Data Searchery demo available here!

http://datasearchery-prototype.research-infrastructures.eu/datasearchery#/search

DATA SEARCHERY