



# **DATA SEARCHERY**

## Preliminary Analysis of Data Sources Interlinking

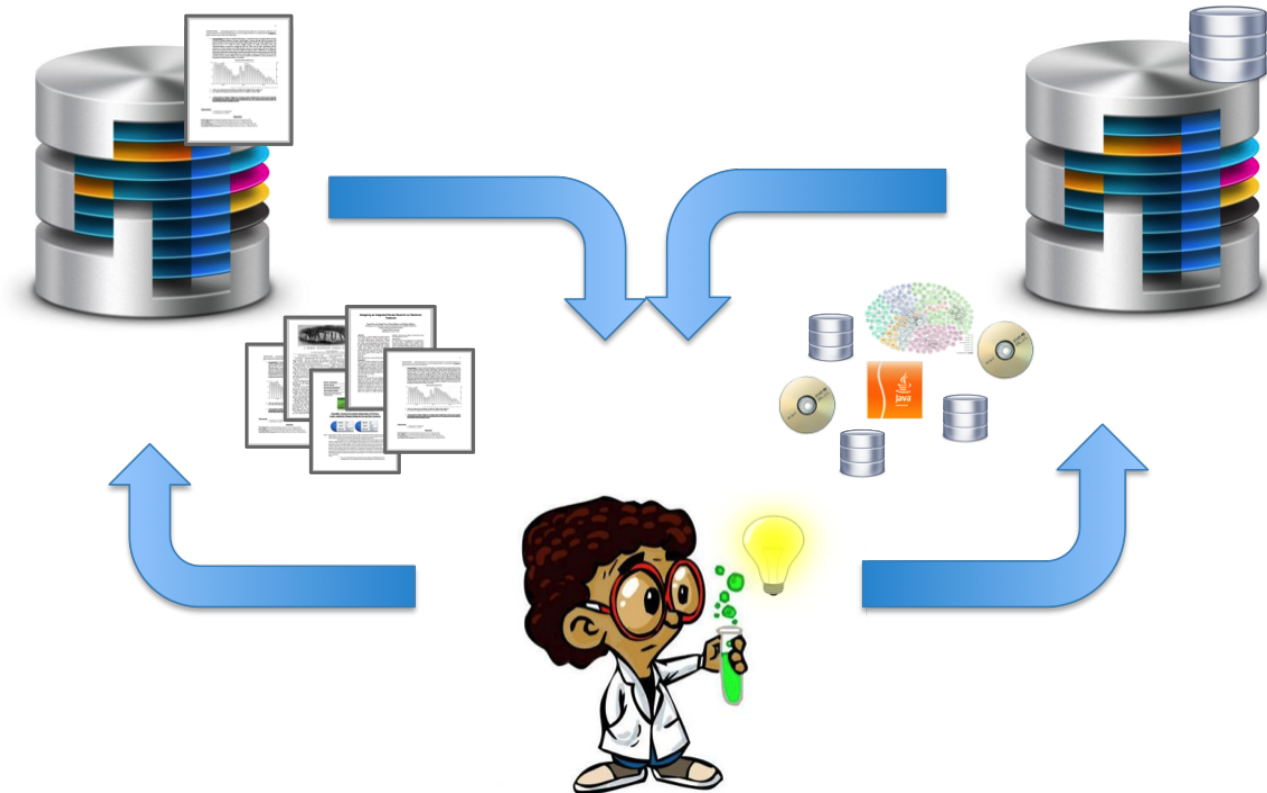
Andrea Mannocci and Paolo Manghi  
ISTI-CNR



# Modern eScience workflow

Research Digital  
Libraries

Research Data Repositories  
@Data Centres

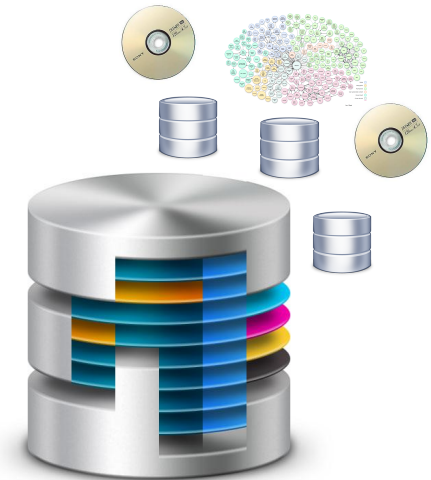
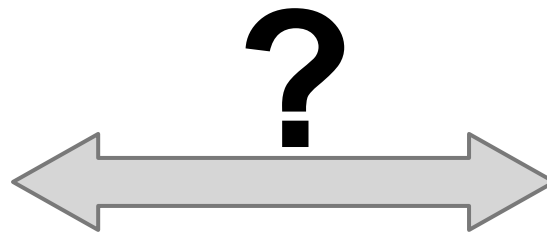


# Modern eScience workflow

*Lack of tools for data-publication interlinking*



Research Digital Libraries



Research Data Repositories

## Benefits:

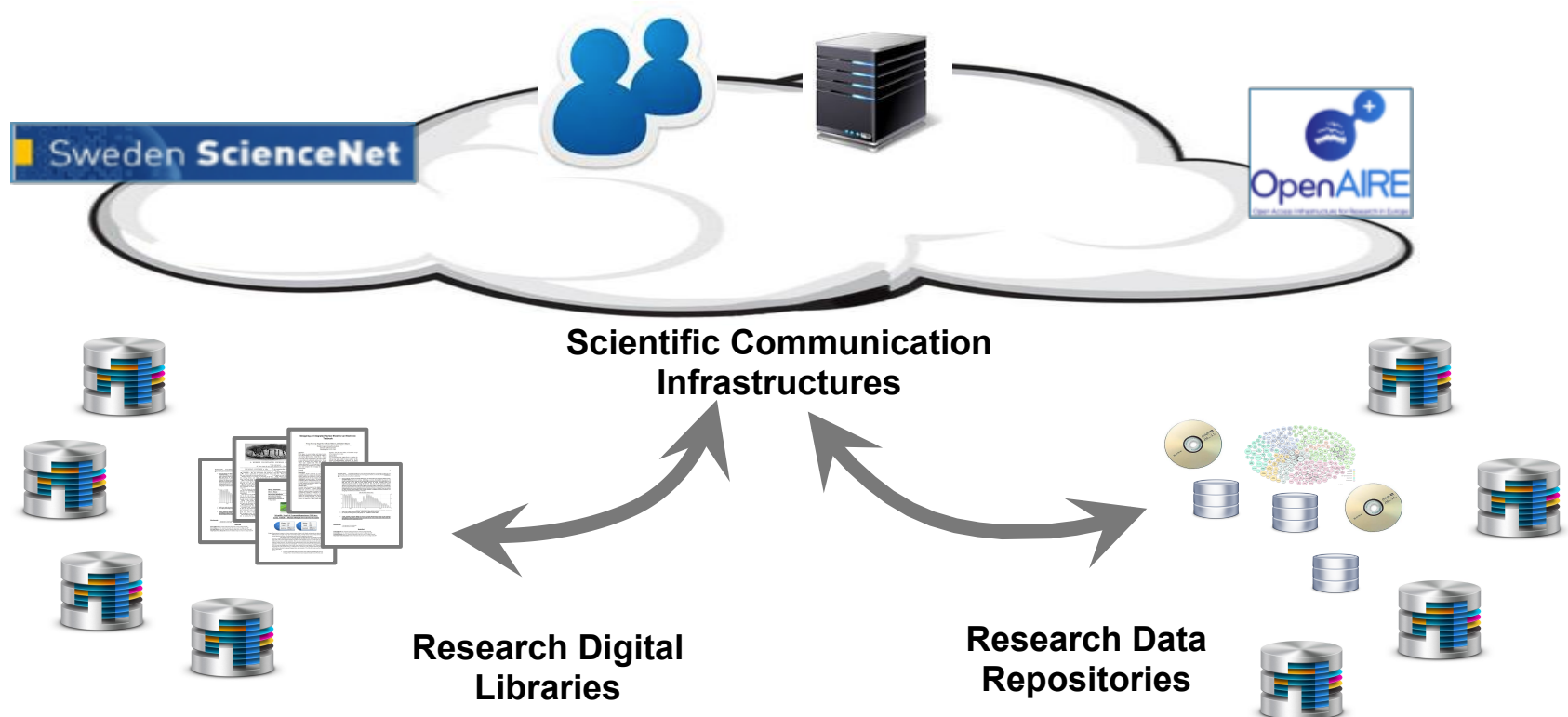
- Foster multidisciplinary research by looking at adherences among distinct disciplines
- Enable better review, understanding, reproduction and re-use of research activities

# 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

Design a tool...

- Light
- Flexible

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

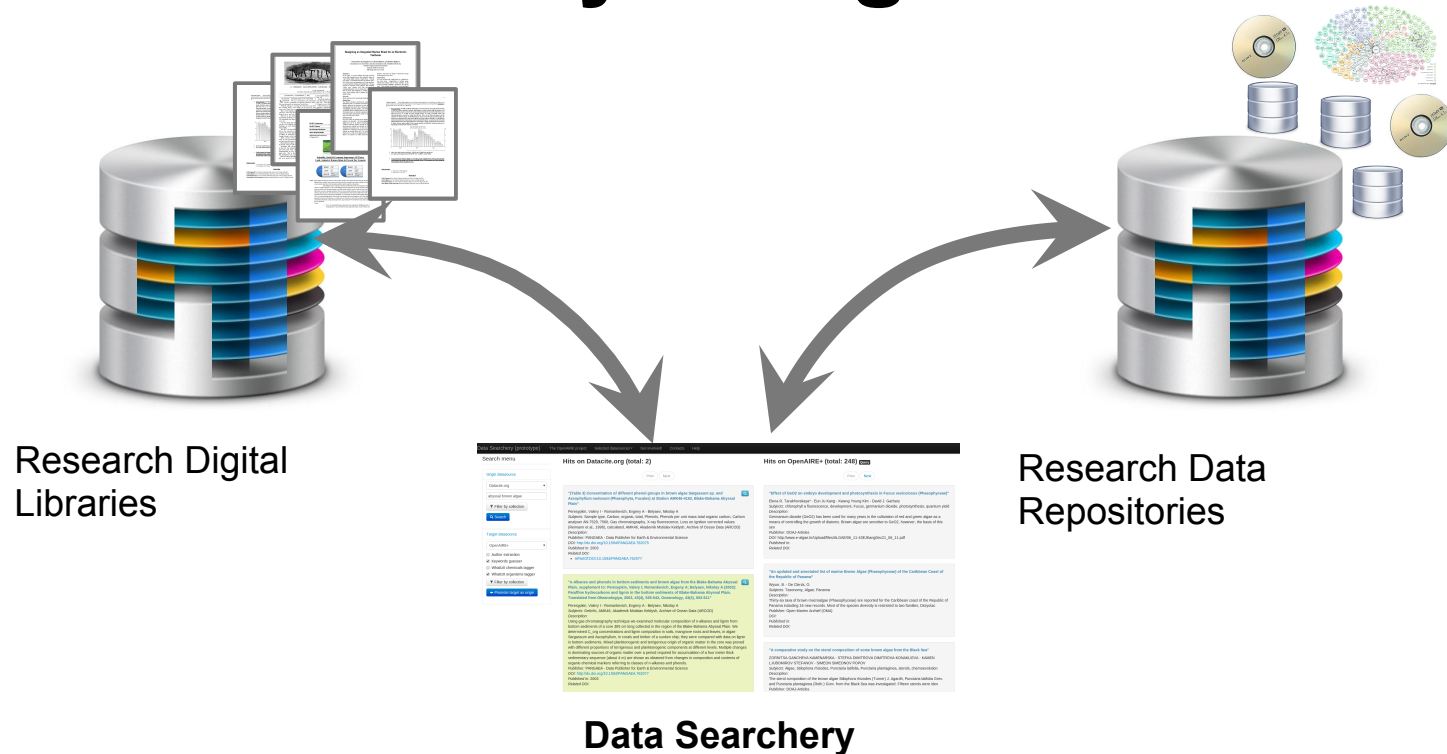
In such a way:

- 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!

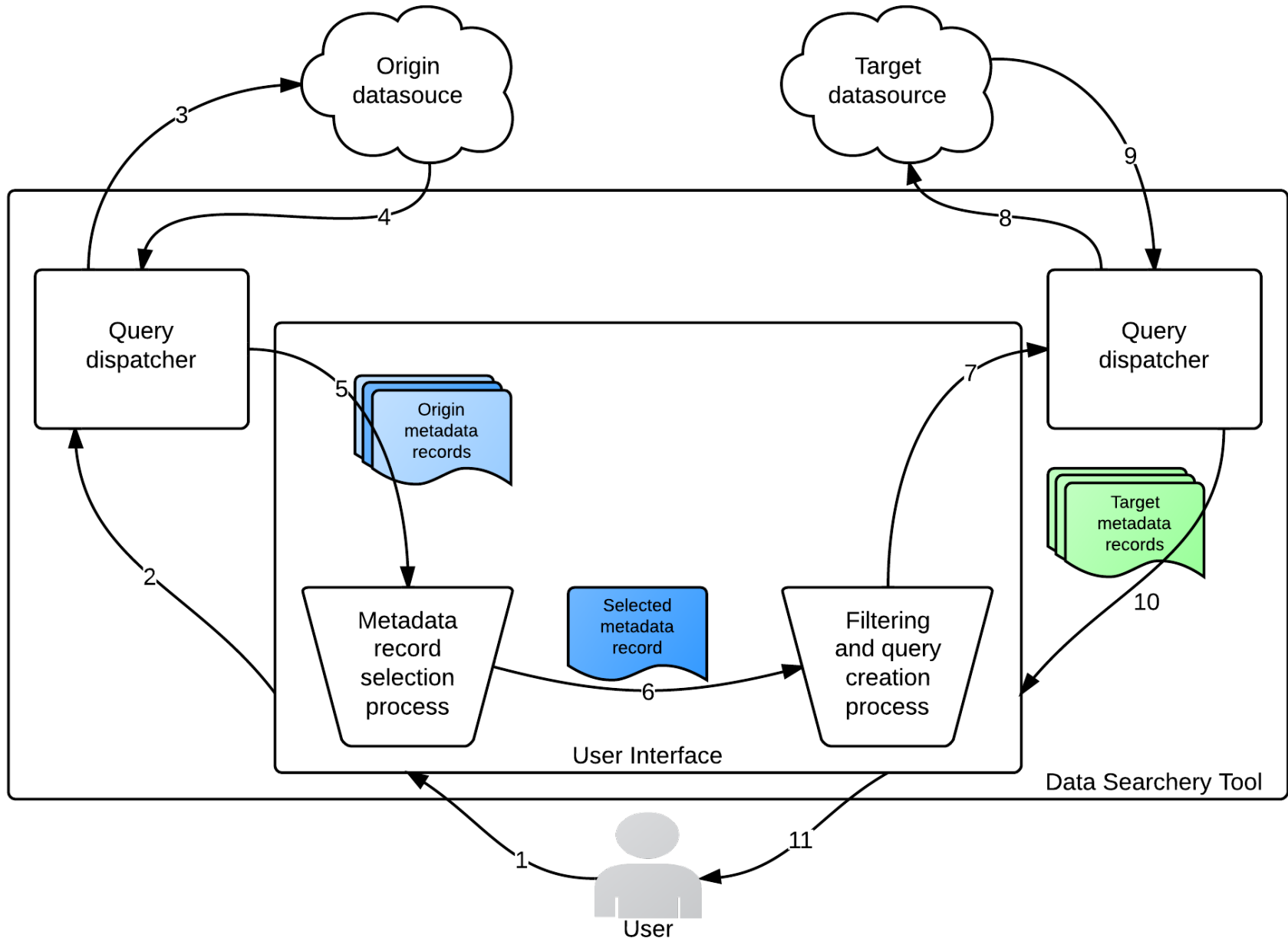
- It can be used as an alternative to SCI, whenever SCIs are not affordable
- It can be integrated to existing SCIs as an additional tool for mining

# 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 at a glance





# Data Searchery

## Main actors in play

### Data Source

- Export of XML-formatted metadata
- Apache Solr 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

Data Searchery {prototype}
The OpenAIRE project
Selected datasources ▼
Get involved!
Contacts
Help

Search menu

Origin datasource

Datacite.org
calcification foraminifer
Filter by collection
Search

Target datasource

OpenAIRE+
Author extraction
Keywords guesser
Whatizit chemicals tagger
Whatizit organisms tagger
Filter by collection
Promote target as origin

Hits on Datacite.org (total: 7)

Prev
Next

**"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 CaCO<sub>3</sub> shells. Yet, CaCO<sub>3</sub> precipitation acidifies the calcifying fluid. Calcification without pH regulation would therefore rapidly create a negative feedback for CaCO<sub>3</sub> precipitation. In unicellular organisms, like foraminifera, an effective mechanism to counteract this acidification could be the externalization of H<sup>+</sup> 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<sup>+</sup> 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](https://doi.org/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: 6) Query

Prev
Next

**"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 CaCO<sub>3</sub> shells. Yet, CaCO<sub>3</sub> 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:* Geochemistry  
Culture experiments with living planktic foraminifers reveal that the ratio of boron to calcium (B/Ca) in *Orbulina universa* increases from 56 to 92 μmol mol<sup>-1</sup> when pH is raised from 7.61 ± 0.02 to 8.67 ± 0.02.  
*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:* Geochemistry  
*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</i>)  
*Publisher:*

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

## 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>



**THANK YOU**

**GRACIAS**  
**ARIGATO**  
**SHUKURIA**  
**JUSPAXAR**  
**DANKSCHEEN**  
**TASHAKKUR ATU**  
**YAQHANYELAY**  
**SUKSAMA**  
**EKHMET**  
**TINGKI**  
**BIYAN**  
**SHUKRIA**  
**GOZAIMASHITA**  
**EFCHARISTO**  
**KOMAPSUMNIDA**  
**MAAKE**  
**GRAZIE**  
**MEHRBANI**  
**PALDIES**  
**BOLZIN**  
**MERCI**