

---

# **Enhancing diffusion of scientific contents: Open data in Open Archives**

Daniela Luzi, Rosa Di Cesare,  
Roberta Ruggieri, Marta Ricci

Institute for Research on Population and Social Policies

# Index

---

■ **Some definitions of data and dataset**

■ **Objectives and methods**

■ **Results:**

- **Providers**
- **Archives**
- **Datasets**

■ **Conclusions**

# Objectives

---

- E-science produces a great amount of data which are shared by a growing number of users;
- Governments and international research institutions are advocating the free availability and preservation of research data, in many cases funding new research data archives;
- Increasing number of publishers require “additional material” to assess the quality of papers;
- Increasing number of Institutional repository are including datasets within their research products



## ANALYSIS AND IDENTIFICATION OF:

- SCIENTIFIC DATA ARCHIVES
- ARCHIVES' MAIN FEATURES
- *AD HOC* CRITERIA FOR THE ANALYSIS
- TEST FOR METODOLOGY

# What are datasets?

## Research Data are:

- “**Facts, numbers, letters, and symbols** that describe an object, idea, condition, situation or other factors”. (National Research Council, Washington D.C.)
- “Any information that can be stored in **digital form**, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc.” (National Science Foundation, 2005)
- “All research output **other than documents** resulting from research activities” (Data Archiving and Networked Services, NL, 2011)

## Dataset is:

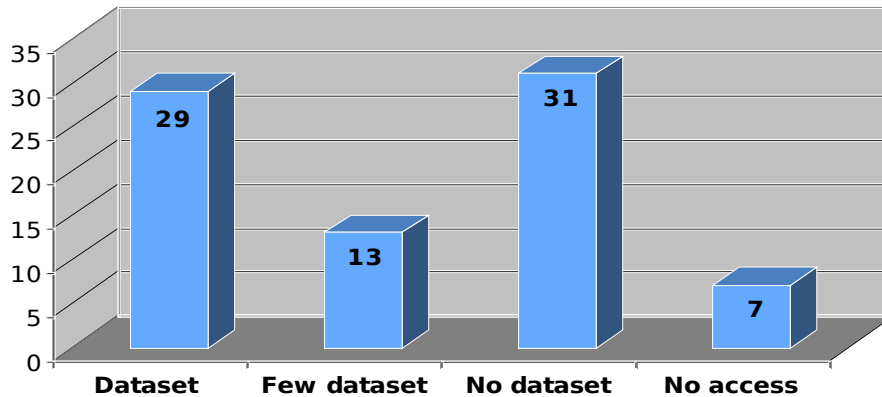
- “A **logically meaningful collection** or grouping of similar or **related data**, usually assembled as a matter of record or for research.” (Online dictionary for Library and Information Science)
- “A **set of files** containing both research data and documentation sufficient to **make data re-use**” (University of Edinburgh)
- “**No-text** scientific and technical information (DOE)

# Methods

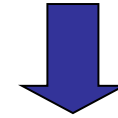
---

- Information source: **OpenDOAR**;
- Selection of the **option “dataset”** from the category “content type”;
- For each identified **archive**:
  - Search for “dataset”;
  - Exclusion of archives with a limited number of datasets (> 5 records);
  - Video, audio and multimedia not considered;
- Identification of **variables** for the analysis;
- Use of some **criteria** adopted by the NSF

# The survey

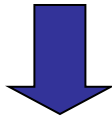


Source of analysis



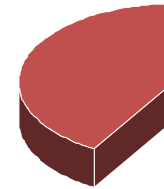
in **OpenDOAR**

Our sample

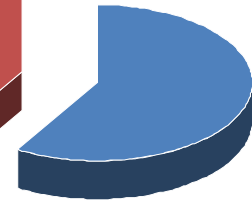


Archives containing datasets

With other  
digital  
objects  
41%



Only  
dataset  
59%



# Who are the Providers?

# What type of Archives?

## Single research Institution

13

Chiba University - Spanish National Research Council - Cambridge University Library and Computing Service - University of Southampton Data Library - University of Edinburgh - Inter America Institute for Global Change Research - International Food Policy Research Institute - University of Minnesota - Monash University Library - University of Delaware Library - University of Hull  
CDS (Centre de Données astronomiques de Strasbourg) - MBLWHOI Library

## Research consortium

7

Mineralogical Society of America - Université du Maine, Le Mans-Laval  
CRL (Center for Research Libraries) - Stichting SURF - Alfred Wegener Institute for Polar and Marine Research (AWI) - Department of Geosciences, University of Arizona - Scripps Institution of Oceanography (SIO)

## Indexing / abstracting service

4

Archaeology Data Service - Scholars Portal OCUL - CBI (National Center for Biotechnology Information) National Library of Medicine (NLM) - EDINA -

## Government

3

Ministério da Saúde-DOE (U.S. Department of Energy) - Deutschen Zentrum für Luft- und Raumfahrt

## Publisher

2

NESCent (National Evolutionary Synthesis Center) - FigShare

## Subject-based Repository

15

American Mineralogist Crystal structure Database - Archaeology Data Service - Crystallography Open Database (COD) - e-Depot Nederlandse Archeologie (eDNA) eCrystals, Southampton - IAI Search - Metropolitan Travel Survey Archive - PubChem - PANGAEA® - RRUFF Project - ShareGeo Open - SIOExplorer  
Digital Library Project - Verkehrsmodelle - Vizier Catalogue Service - Woods Hole Open Access Server

## Institutional Repository

7

Chiba University's Repository for Access To Outcomes from Research (CURATOR) Digital.CSIC - DSpace @ Cambridge - Edinburgh DataShare - Monash University ARROW Repository - University of Delaware Library Institutional Repository - University of Hull Institutional Repository

## Directory

6

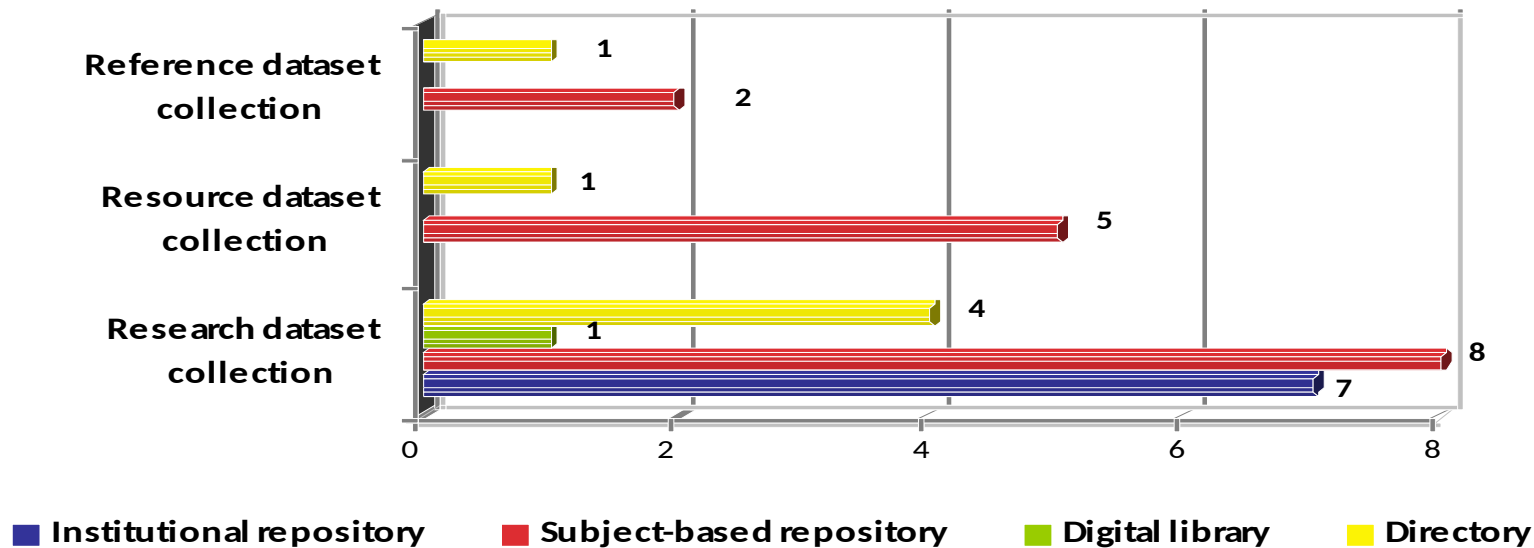
Biblioteca Virtual em Saúde - Dryad - FigShare - IFPRI Publications (International Food Policy Research Institute Publications) - OSTI - OZone

## Digital library

1

DSLA (Digital South Asia Library)

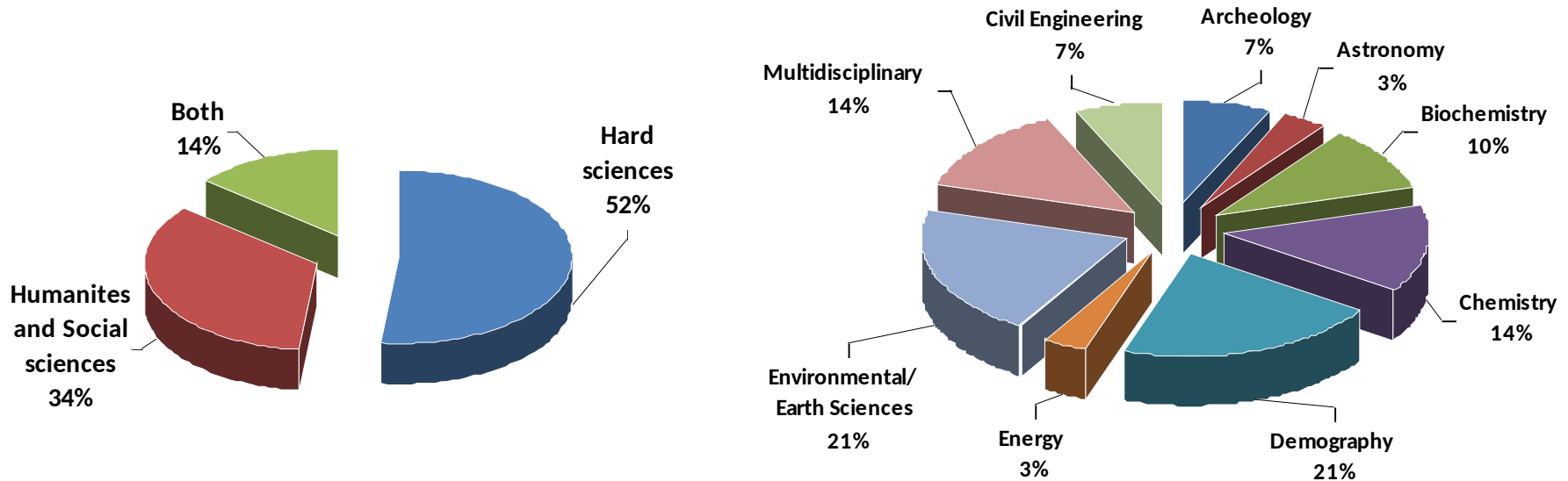
# Functional categories of digital data collections



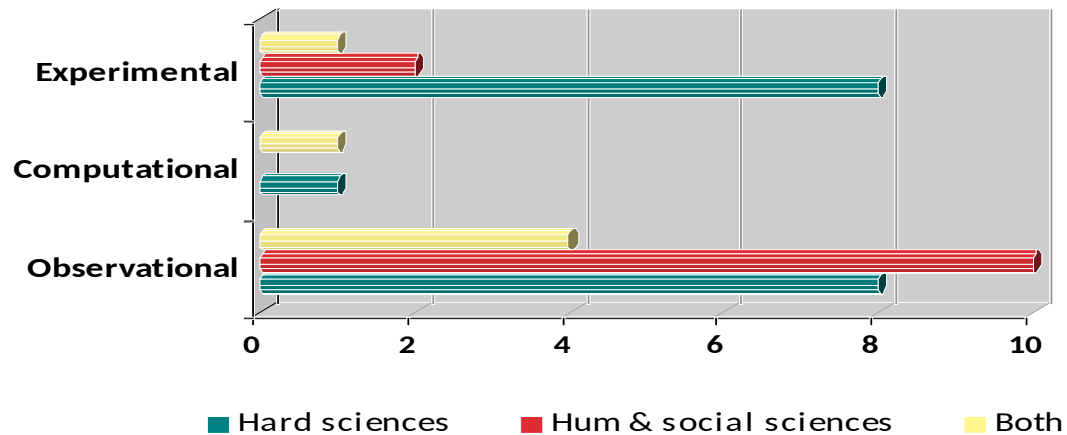
- **Research data collections:** *focused research project, serves a specific group, small budget;*
- **Resource data collections:** *focused on a community, serves a single science, may develop community-level standards, direct funding from agencies;*
- **Reference data collections:** *serves a large set of scientific community, uses well-established standards, multiple funding sources, foresees indefinite maintenance. (National Science Foundation, 2005)*



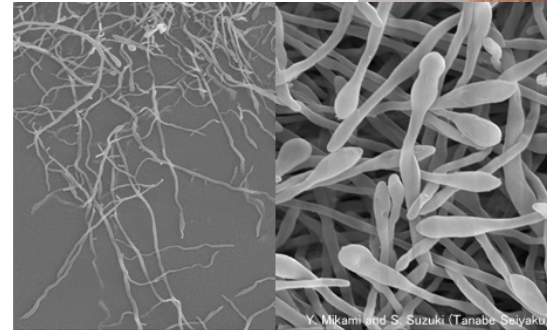
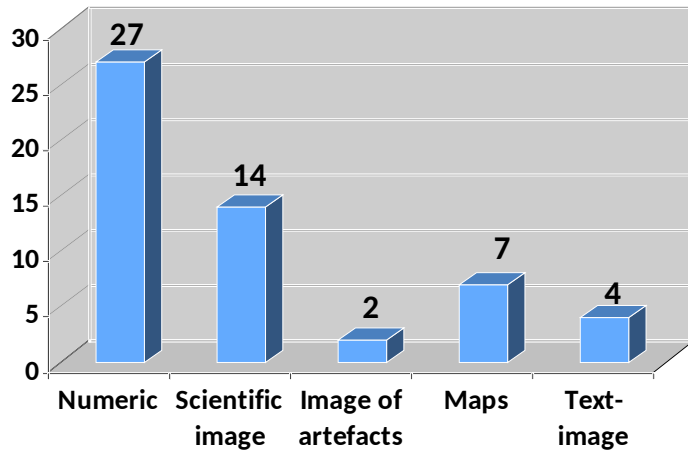
# Which science area?



## Data origin



# Types of dataset



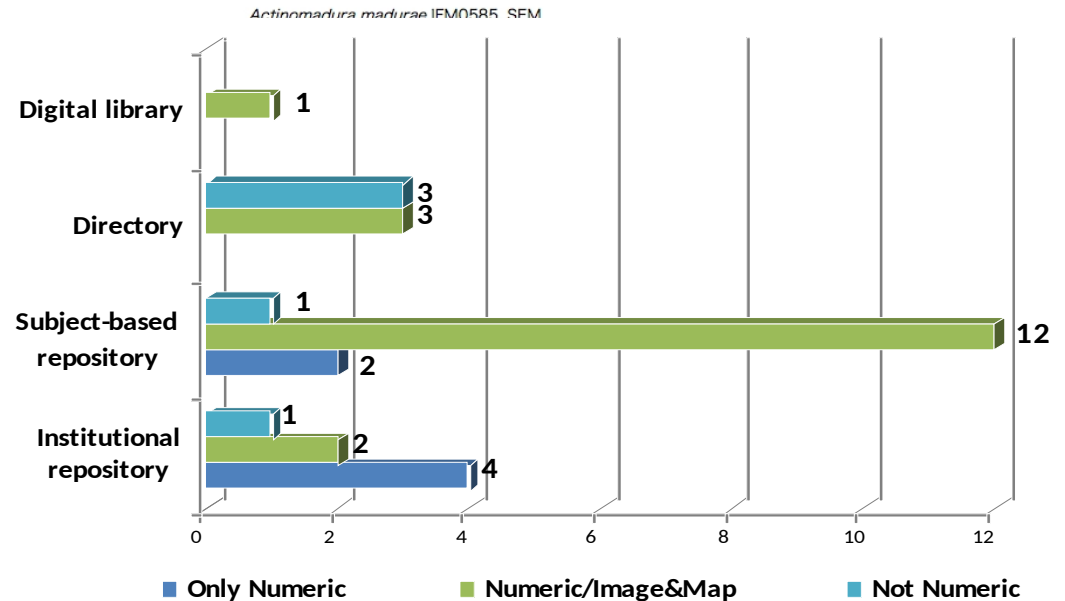
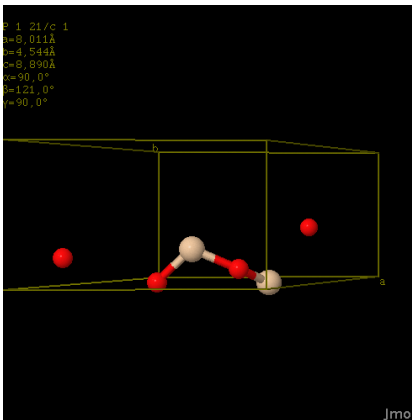
## American Mineralogist Crystal Structure Database

4 matching records for this search.

### Cristobalite II

Dera P, Lazarz J D, Prapakempa V B, Barkley M, Downs R T  
 Physics and Chemistry of Minerals 38 (2011) Online-first  
 New insights into the high-pressure polymorphism of SiO<sub>2</sub> cristobalite  
 Note: P = 3.5 GPa, data transformed from Dove et al (2000)  
 \_database\_code\_amcd 0018348  
 8.011 4.544 8.890 90 121.0 90 p2\_1/c  
 atom x y z Uiso  
 Si1 .6274 .2661 .0461 .018  
 Si2 .8660 .0221 .7189 .018  
 O1 .8152 .1255 .8517 .028  
 O2 .6944 .4046 .0275 .028  
 O3 .4637 .0262 .7952 .028  
 O4 .0696 .1948 .7604 .028

[Download AMC data \(View Text File\)](#)  
[Download CIF data \(View Text File\)](#)  
[Download diffraction data \(View Text File\)](#)  
[View Jmol 3-D Structure](#)



# Dataset format

## Formats:

- Flat files (.txt, ascii, .csv)
- Word processor (.doc, .pdf)
- Image (.tiff, .jpeg, gif, .jmol)
- Spreadsheet (.xls)
- Statistical analysis (SPSS)



















Generally available in data package (.zip, .tar)

## ReadMe files:

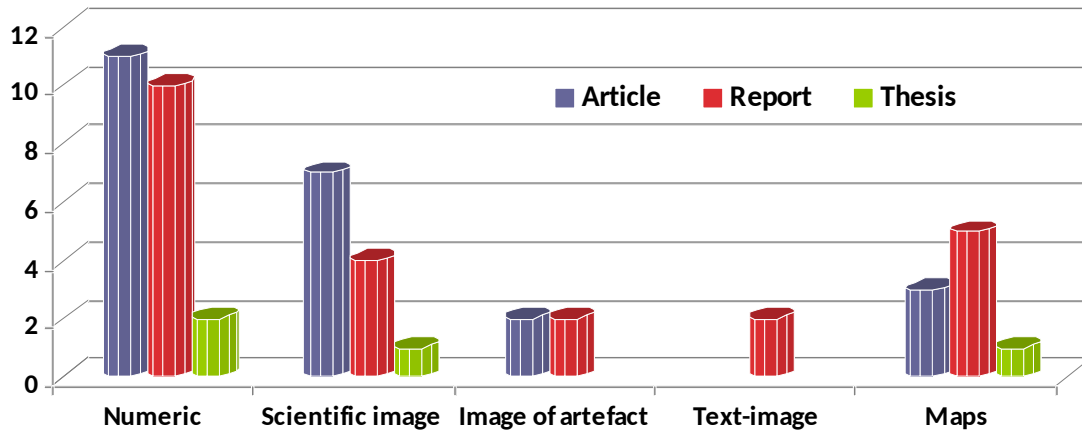
- frequent in research data collection,
- especially in IR (57%) and Directories (67%)

## Specific format for data exchange:

- Especially in resources and reference data collections
- Used only in subject-based repositories (53%)

Files	Description	Dimensione	Formato	Mostra	Description
<a href="#">IST_data.csv</a>	IST semi-colon delimited data file	5.531Mb	Text file	 Download 	IST semi-colon delimited data file
<a href="#">IST_data.txt</a>	IST unicode data file	11.10Mb	Text file	 Download 	IST unicode data file
<a href="#">IST_database_paper.doc</a>	IST database paper in MS Word format	89.5Kb	Microsoft Word	 Download 	IST database paper in MS Word format
<a href="#">IST_database_paper.pdf</a>	IST database paper in PDF format	86.40Kb	PDF	 Download 	IST database paper in PDF format
<a href="#">IST_variables.csv</a>	IST variables in csv format	5.985Kb	Text file	 Download 	IST variables in csv format
<a href="#">IST_variables.txt</a>	IST variables in unicode format	12.11Kb	Text file	 Download 	IST variables in unicode format
<a href="#">IST_variables.pdf</a>	IST variables in PDF format	64.79Kb	PDF	 Download 	IST variables in PDF format
<a href="#">IST data suppl 1.csv</a>	IST variables in csv format	4.604Mb	Text file	 Download 	IST variables in csv format
<a href="#">IST database open access paper.pdf</a>	IST database paper in PDF format	178.7Kb	PDF	 Download 	IST database paper in PDF format

# Datasets linked with ...



Only a few datasets have no association with other items (27.6%)

Other types of associations:

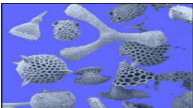
- description of the collection
- description of the project

**CORED SEDIMENTS**

[Cored Sediment and Microfossil Collection](#) [Print-Friendly Version](#)

**Core Collection**

Deep-sea sediment cores are vital to our understanding of the past and present oceans. They record the geological history of the ocean basins, providing evidence for changing climates, emerging environments, evolving biota, and dramatic events that have altered the course of earth history.



The SIO collection contains nearly 6,600 cores (15,000 refrigerated core sections) collected using gravity, piston, trigger, vibra- and box-coring techniques. It is the largest collection in the U.S. (outside that of the Ocean Drilling Program) of sediments from the Pacific Ocean and also contains extensive material from the other major ocean basins.

*Miocene radiolarians. Photo: R. Norris.*

**Microfossil Collection**

Microfossils are the skeletal remains of marine organisms that either floated in the water column or lived on the sea floor. The Collections contain the raw samples, prepared microscope slides, and field notes from pioneering paleontologists who were the first to recognize and implement the use of marine microfossil remains for dating and correlating sediments. These include, for example, the very extensive Riedel/Sanfilippo radiolarian collection, M. N. Bramlette's calcareous nannofossil preparations, Fred Phleger's and Frances Parker's foraminifer collections (in part) and Patricia Doyle's microfossil fish teeth (ichthyoliths) collection.

The Collections support not only SIO faculty and student research, but also that of scientists from other domestic and non-U.S. institutions. Studies relate to virtually all fields of earth science including paleoceanography, paleoclimatology, stratigraphy, paleontology, geochemistry, geophysics, mineralogy and tectonics. Collections materials provide a viable resource for the education of graduate, undergraduate and K-12 students.

**Staff**  
 Dr. Richard (Dick) Norris, Curator. Email [norris@ucsd.edu](mailto:norris@ucsd.edu)  
 Dr. Annika Sanfilippo, Curatorial Advisor. Email [annika@ucsd.edu](mailto:annika@ucsd.edu)  
 Mr. Warren L. Smith, Collections Manager. Email [wsmith@ucsd.edu](mailto:wsmith@ucsd.edu)

**Campus Location**  
 Deep Sea Drilling Building-West, Rooms 55 and 91.

## Meols: The Archaeology of the North Wirral Coast

David Griffiths, Robert Philpott, Geoff Egan, 2011

[Introduction](#)  
[Overview](#)  
[Query](#)  
[Downloads](#)

Data copyright © National Museums Liverpool unless otherwise stated



### Primary contact

Dr Robert Philpott  
 Head of Archaeology  
 Museum of Liverpool  
 Dock Traffic Office  
 Albert Dock  
 Liverpool  
 L3 4AX  
 England  
 Tel: 0151 4784337

[Send e-mail enquiry](#)

### Introduction

The aim of the project was to catalogue all the finds known to have been recovered from the North Wirral shore in the Meols area, to interpret these in the light of modern scholarship and to publish the results. The majority of the finds were recovered during the 19th century from the eroding coastline at Meols. Many finds were published in the 19th century, along with topographic observations.

The dataset consists of extant finds which are now in museum collections in a few cases private ownership, and these are presented in a database with digital photographic or scanned images.

Well-recorded non-extant finds and clay tobacco pipes are listed in the database but are not illustrated in the dataset submitted here. Numismata are not submitted here. These categories and interpretative text can be found in the Meols monograph published in 2007 (Griffiths, Philpott and Egan 2007).

### References

Griffiths D., Philpott R. A. and Egan G. 2007 *Meols, The Archaeology of the North Wirral Coast. Discoveries and observations in the 19th and 20th centuries with a catalogue of collections*, Oxford University School of Archaeology Monograph Series 68, Oxford.



Wealth Creation, 5-6 December 2011



# Conclusions

*Our sample captured a variety of cases:*

- Providers: Research and **Governmental** institutions, **Publishers**, Data services;
  - Collaboration: single institution, **consortium**
- Archives: IR, **Subject-based repositories**, **Directories**, **Digital libraries**;
  - Only dataset: **concentrated in Subject-based repositories**
- Discipline: In theory all, tend to be **sub-disciplinary fields**

- Data origin: **prevalence of observational dataset**
- Data collection categories:
  - IRs: **only research collection**,
  - Subject based repositories and directories: **also resource and reference data collections**

- Types of dataset:
  - **numeric data represented also with images and maps: especially in Subject-based repositories and Directories**
  - **specific exchange format: only within Subject-based repositories**

# Discussion

---

- What are the most suitable types of archives for dataset?
- Are IRs progressively inserting dataset along with other digital objects?
- Can we think of developing IR exclusively devoted to dataset produced within the institution?
- Can digital libraries contribute to diffuse re-usable dataset?
- Are there interconnections between dataset and sub-disciplinary fields?
- What makes dataset reusable?
- What about preservation?