Data Management for Environmental Informatics:
An Irish Research Perspective

Peter Mooney and Adam Winstanley

Contact Information
Dr. Peter Mooney

Environmental Research Center (ERC),

Environmental Protection Agency,

Richview,

Clonskeagh,

Dublin 14.

Ireland.

Ph: +353 (1) 268 0100

National Center for Geocomputation,

John Hume Building,

National University of Ireland,

Maynooth,

Co. Kildare.

Ireland.

Email: peter.mooney@nuim.ie

Part of RESEARCH DEPT in the Environmental Protection Agency (EPA)

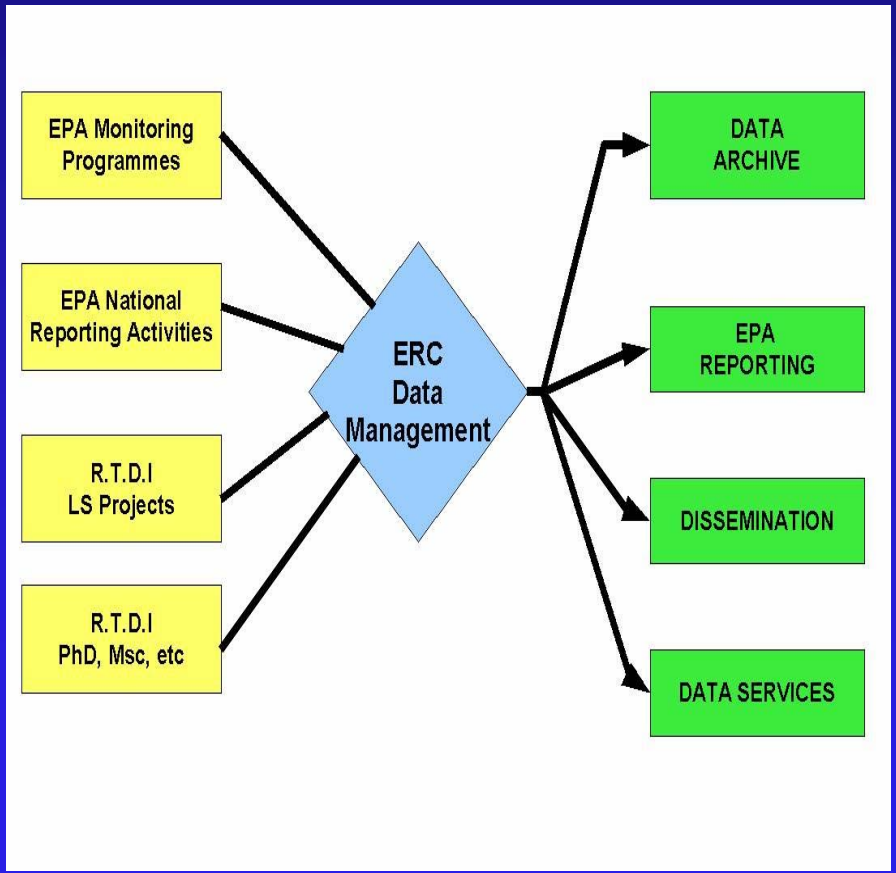
- €50 Million investment (2000 – 2006)
- Structured approach to Irish Environmental Research
- ERC Working Areas:
  - Research Data Management,
  - Climate Change,
  - Transboundary Air Pollution,
  - Strategic Environmental Assessment (SEA),
  - Water Framework Directive (WFD)

What are Environmental Data?

- "**Any measurements or information that describe environmental processes, location, or conditions; ecological or health effects and consequences; or the performance of environmental technology**"

- Environmental data include:
  - information collected directly from measurements,
  - produced from models,
  - compiled from sources like databases or the literature
  - Licence information,
  - Reporting obligations

Our Principal Role is Data Management and Informatics for EPA Research



- Providing a focal point for collection of data from our funded projects in Ireland
- Includes special data services
- Pro-active approach to collaborative data exchange and data archive

Considerable Data Volumes are Generated By Research Programmes

Research Programmes

MSc, PhD, PostDoc
Small, Med, Large Scale

Reports
Scholarly Publications
RAW Data
Derived Data

EPA

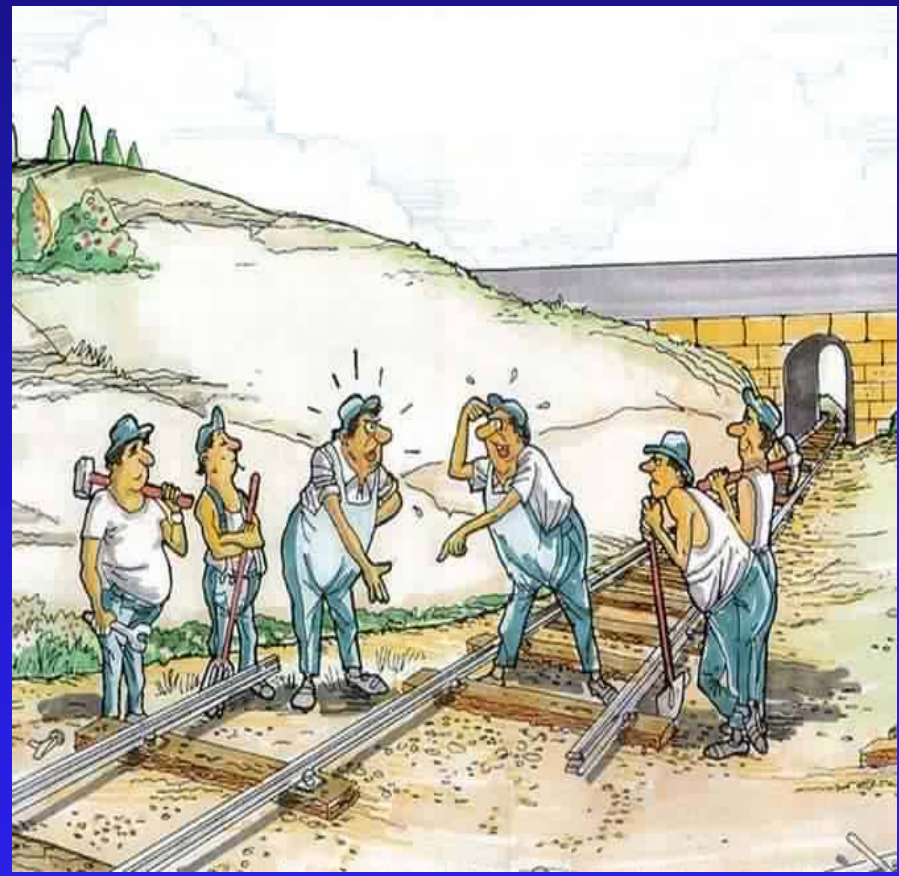
ERC
Data Archive

Data

Environmental Valuable Assets

Currently No Research Data Repository Infrastructure In Ireland

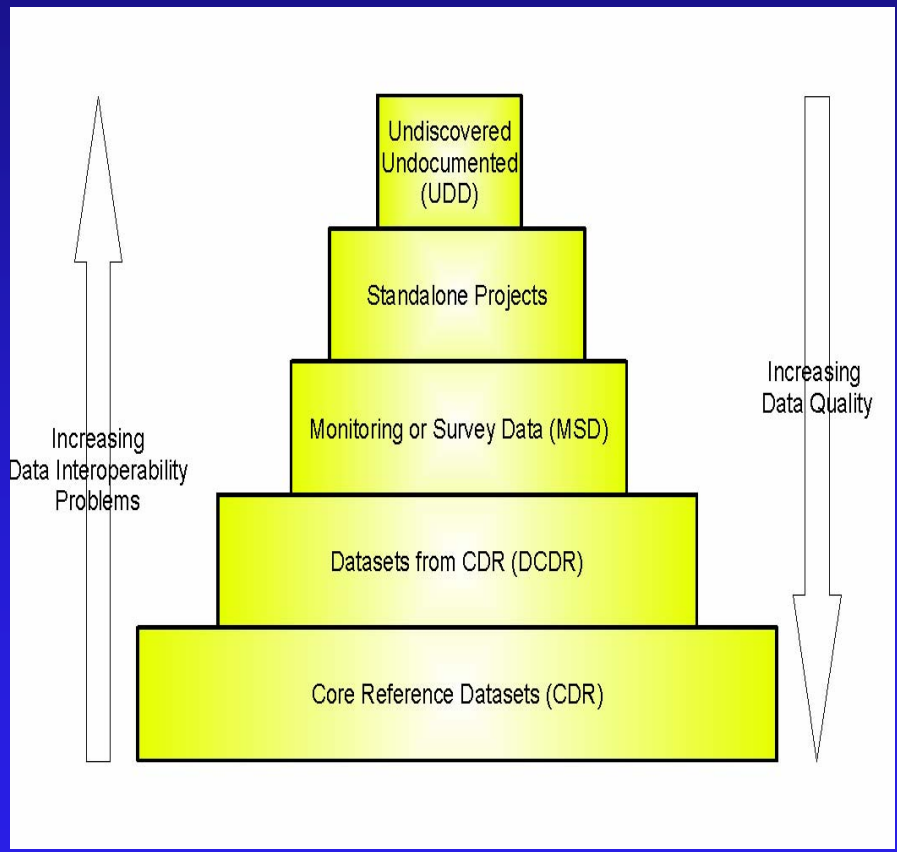
- Irish Physical Science research funded by many different agencies
- Researchers working in isolation – often focussing on "*grant-getting-approaches*" **(Eric Kihn)**
- Indicators of success is still traditional peer review + ability to attract funding
- Data is NOT REWARDED



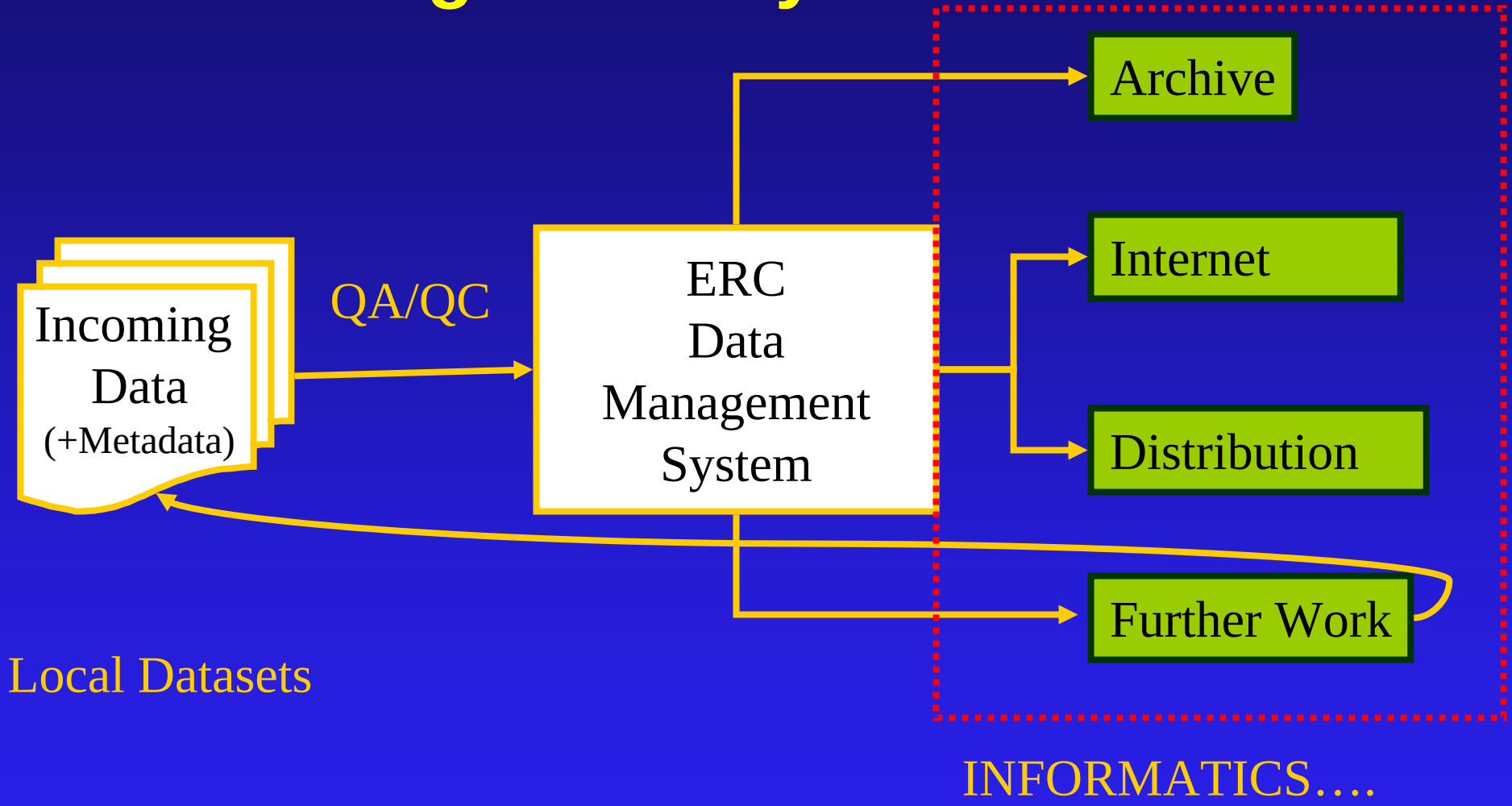
Lack of Coordination

All Data Are Created Equal:
Some Are Managed Better Than Others

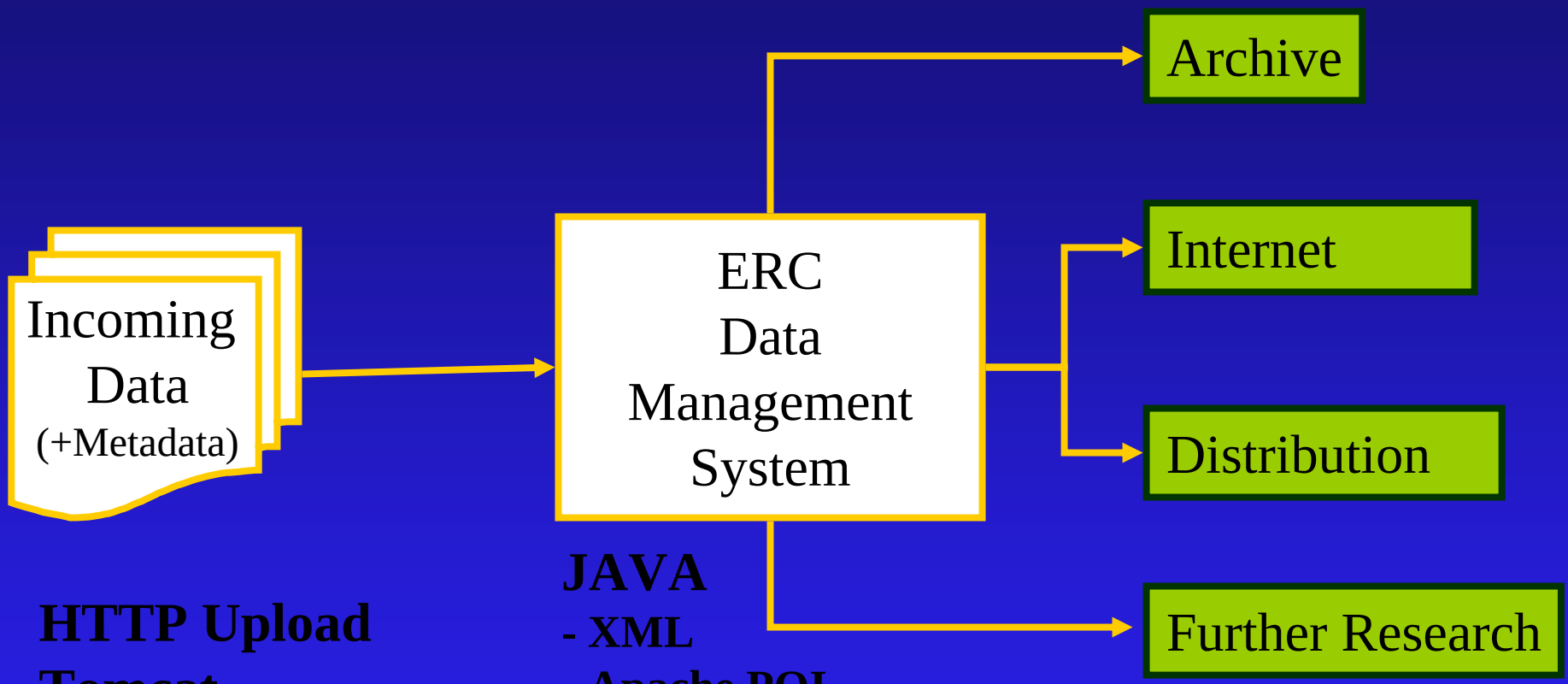
- Large Scale National Level projects are usually the best for Interoperability and Data Quality
- Small “localised” projects – many interoperability problems for a variety of reasons



Description of our Data Management System

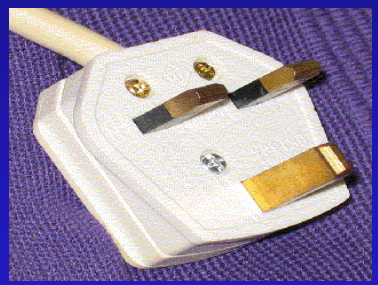


The ERC Data Management System uses Several Different Software Tools



Interoperability problems occur when exchanging services between different system specifications

Service Consumer



Service Provider

User (Consumer)

System Type “T1”

Consumes Formats X, Y Z

Serves Formats P, Q, R, S, & T

Server (Service Provider)

System Type “S1”

Interoperability is encountered in several different working contexts

- Problems due to the types of computer hardware used
- Problems due to the types of computer operating system used
- Problems due to the types of measurement instrumentation

HARDWARE

- Data Exchange – systems do not understand each others formats
- Semantic Problems in Data Exchange
- IPR or Copyright issues in data exhange or use

SOFTWARE or HUMAN

Most Environmental Data Undergo QA/QC processes before general release

- **Data Outlier Filtering**
  - System Outliers vrs Suspicious Outliers
- **Range Rationality Checking**
  - Parameters exceeding the range of Sensors
  - Values outside the phsysical restrictions of the environment

Measurement/Calculation

- **Data Type Checking**
  - Numerical Data Types checked for consistency
- **Temporal Consistency Checking**
  - ISO 8601 YYYY-MM-DDThh:mm:ss

Storage/Structure

Our Funded Researchers Must Submit Final Reports and All Raw Data

Reports, Papers, etc

Submit

EPA

ERC
Data Archive

DATA
ISSUES

Generating
Data

Reads
DataMgmt
Guidance

Data
QA/QC

Submit
Upload

Start

Project Timeline

End

Revision of the Framework for Data Capture From Research Projects

1. More “Pro-Active Engagements” with the Research community much earlier in the project timeline

2. Researchers to complete a “Data Management Plan”

3. Explore incentives to:
   - Increase Researcher interest in Data Management
   - Make more metadata public

We Are Developing a More Pro-Active Framework for Data Capture

Reports, Papers, etc

Submit

EPA

ERC
Data Archive

Submit
Upload

Generating
Data

DM
PLAN
Accepted

IMPLEMENTS
DataMgmt
PLAN

Data
QA/QC

Start

Project Timeline

End

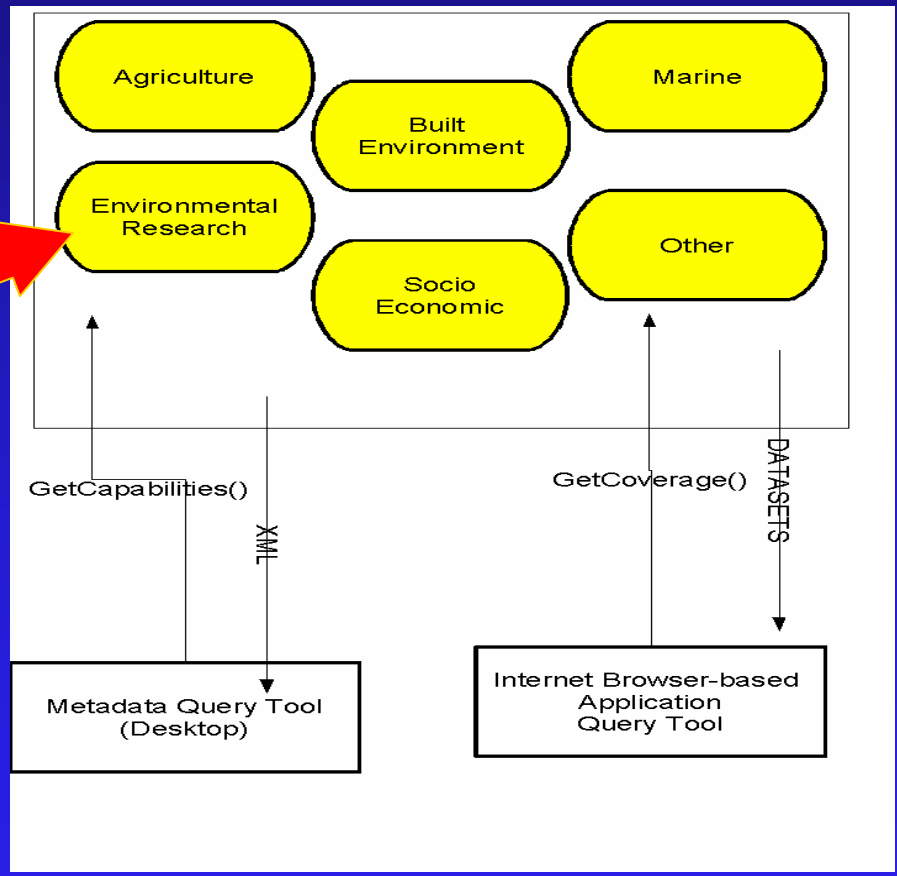
Data Providers (Researchers) still retain a high degree of autonomy

- Researchers are not bound to a ONE-FORMAT-FITS-ALL policy
- Good data management is fostered in the project from the earliest point
- As INSPIRE outlines – data is managed ***as close to the source as is appropriate***

Use of OGC Web Services allows development of “Joined-Up-Services”

- Each funding organisation drives their own data management strategies

- To client – they see Joined-Up-Services
- They have choice of tools
- No expert knowledge needed



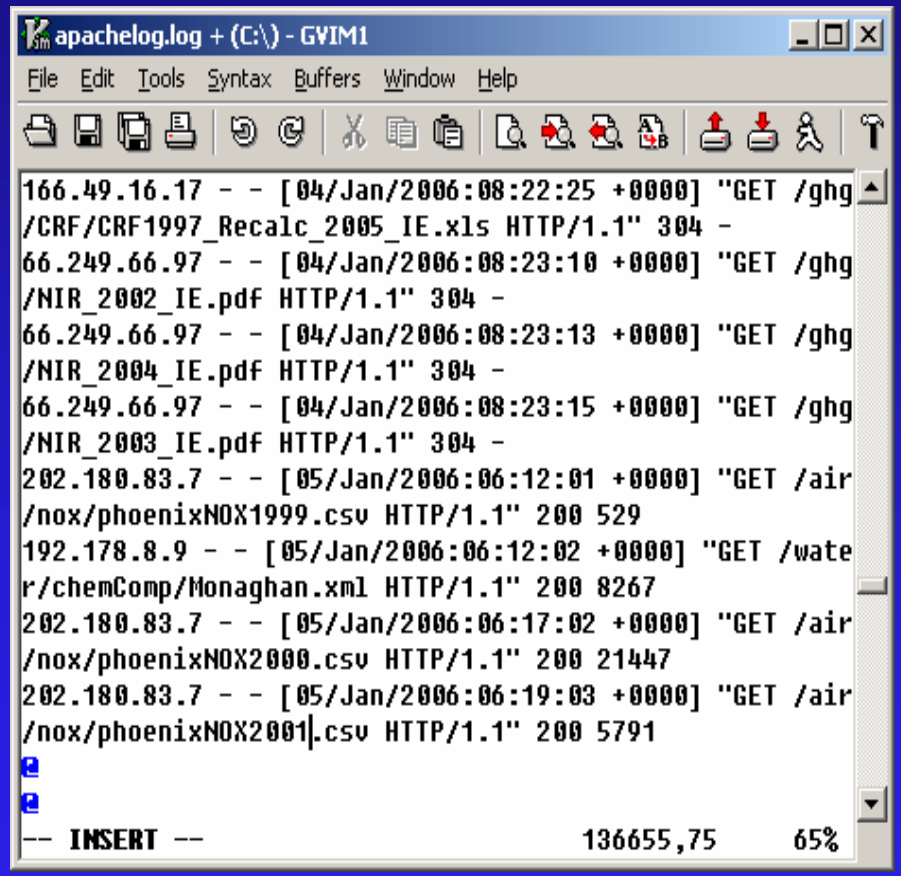
Web Coverage Service Example

OGC Services sees traditional HTML-website data distribution diminishing

- Difficult to maintain currency and consistency of data archives with traditional HTML-based website approach
- OGC Services approach means multiple points of entry and multiple query options to ONE DATASET in ONE LOCATION
- "**Clip-It, Zip-It, Ship-It**" Data Exchange MUST STOP

Provide Feedback to Data Providers on Web-Server Statistics

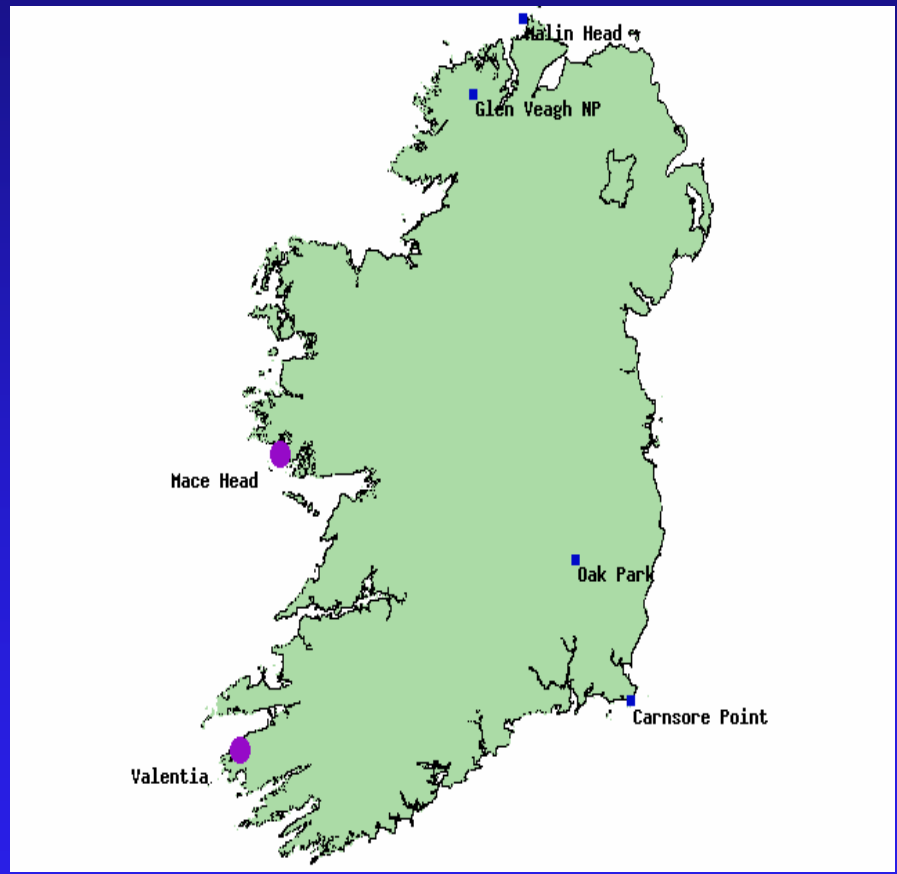
- Encourage data providers by production of frequent data access statistics
- Stats such as
  - Total Data Downloaded
  - Most Popular Datasets
  - Most Viewed Metadata
- Some form of reward mechanism required



Other Issues Arising From This Work

Good Data Management Allows Design of Useful Informatics Solutions

- Transboundary Air Pollution Monitoring
- All stations measure (CO, SO2, O3, Nox) – in XML
- Uploaded to server hourly
- Other International Researchers then download into Air Quality Models



The older (temporally) the Environmental Data is the better

- Often older Envir. Data comes from periods not effected by current changes
- Analysis of the impact of current environmental pressures
- *Example*: Key for WFD Baselines for many water species

“Grey and Dusty” Publication Room – How Do We Search? Spatial Queries?

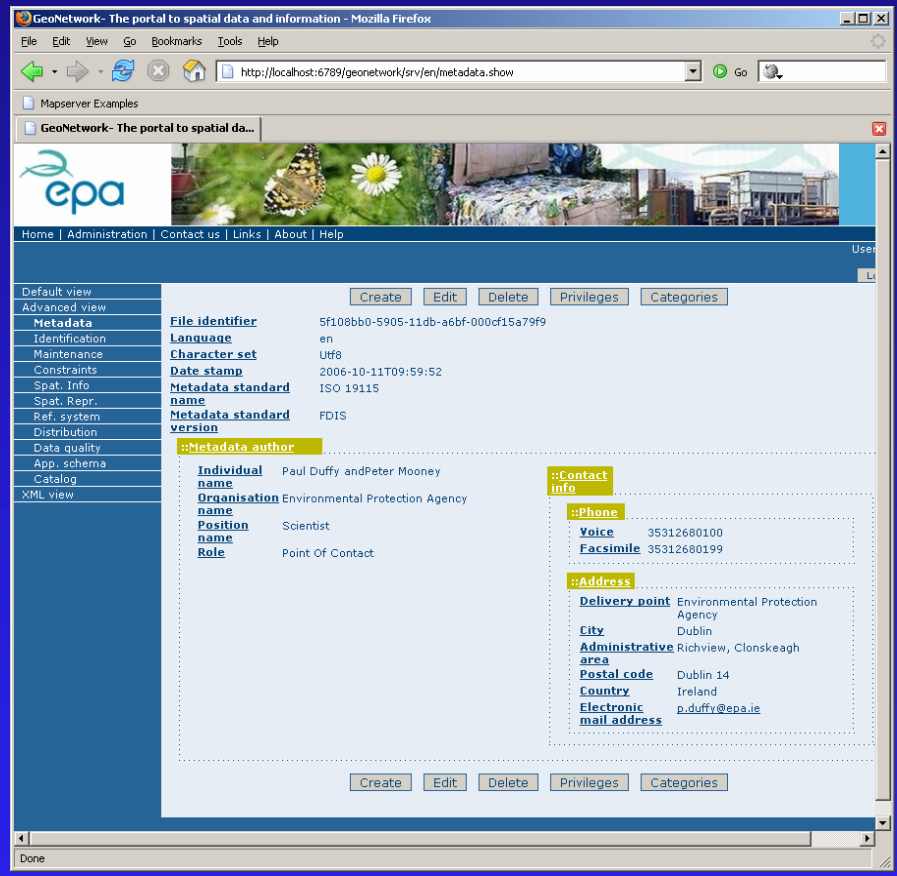
- Vast potential if this “*paper archive*” is brought to digital life”
- Create Searchable Metadata
- Small-scale project with significant results

Data Resources Should Not Be Limited to Standard Notions of "Data"

- The amount of data about the environment far exceeds that captured in traditional data paradigms
- *M. Craglia (JRC, 2005)* – "Think of cataloguing models, multimedia, and services themselves"
- Large amounts of "data" and "information" not yet catalogued or geocoded

GeoNetwork – web based metadata catalogue with OGC compliance

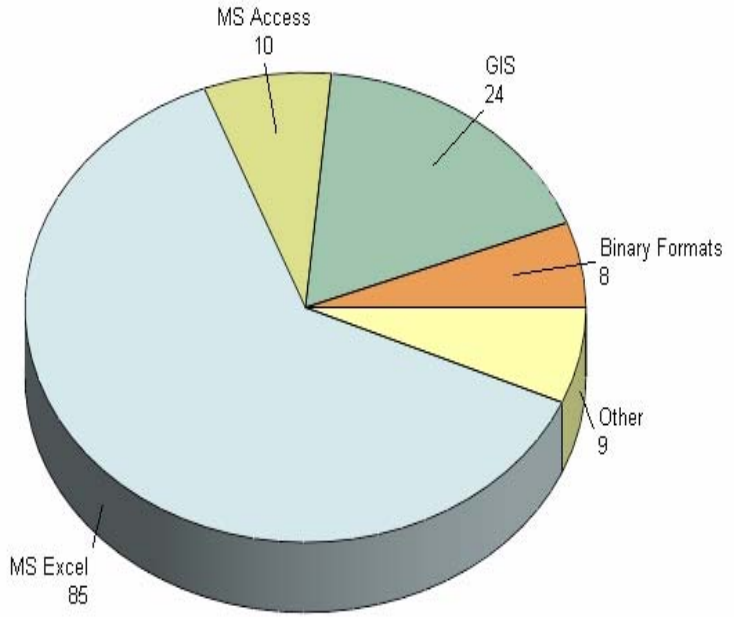
- Free and Open Source Catalog Application
- Metadata Editing and Search
- Integrated Web Map Views
- Full ISO 19115 implemented
- Community Maintenance – More Secure



MS Excel remains a popular choice of software format with researchers

- **Advantage:** Excel offers non-IT specialists:
  - an easy to use package
  - data collection, visualisation,
  - analysis, distribution

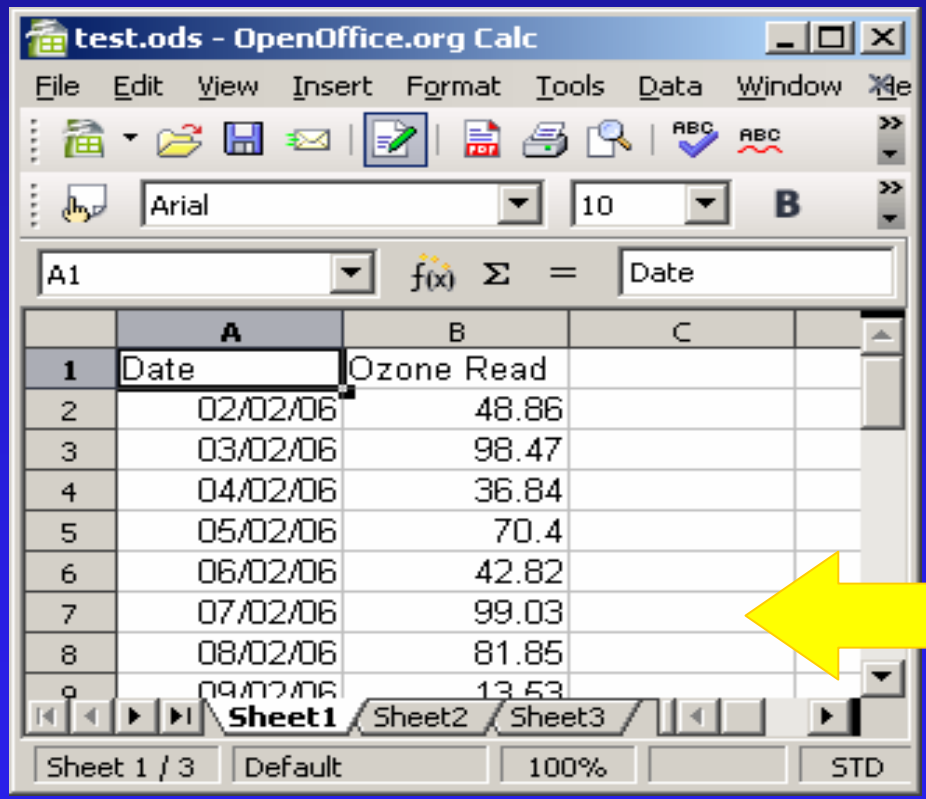
- **Disadvantage:**
  - Poor Data Interoperability
  - Difficult to automate data extraction with 3G languages

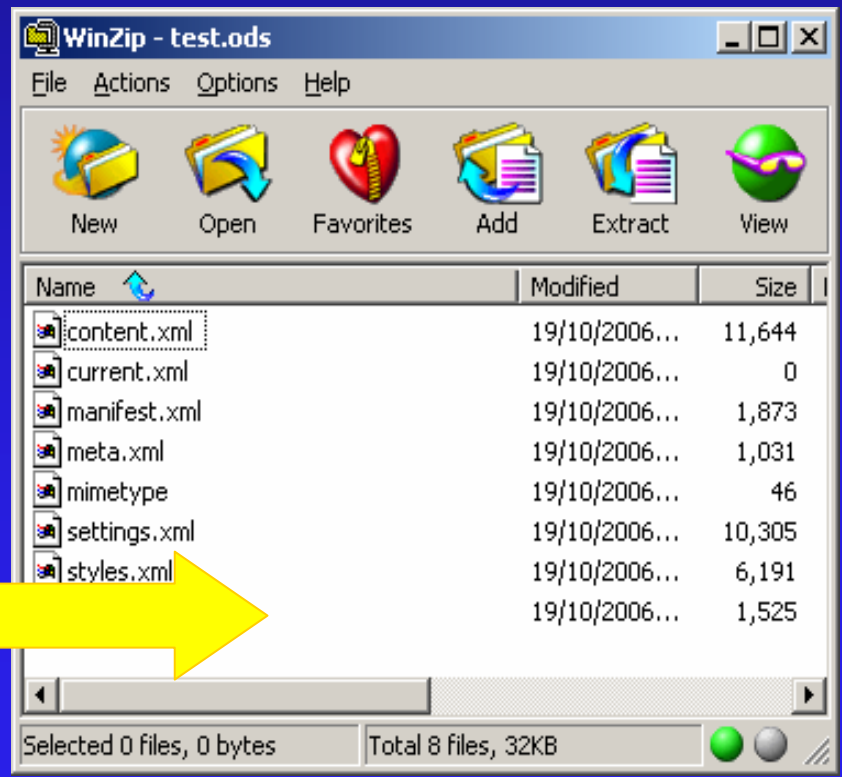


136 PhD Level Projects

Encourage use of Open Document Formats over Closed Proprietary

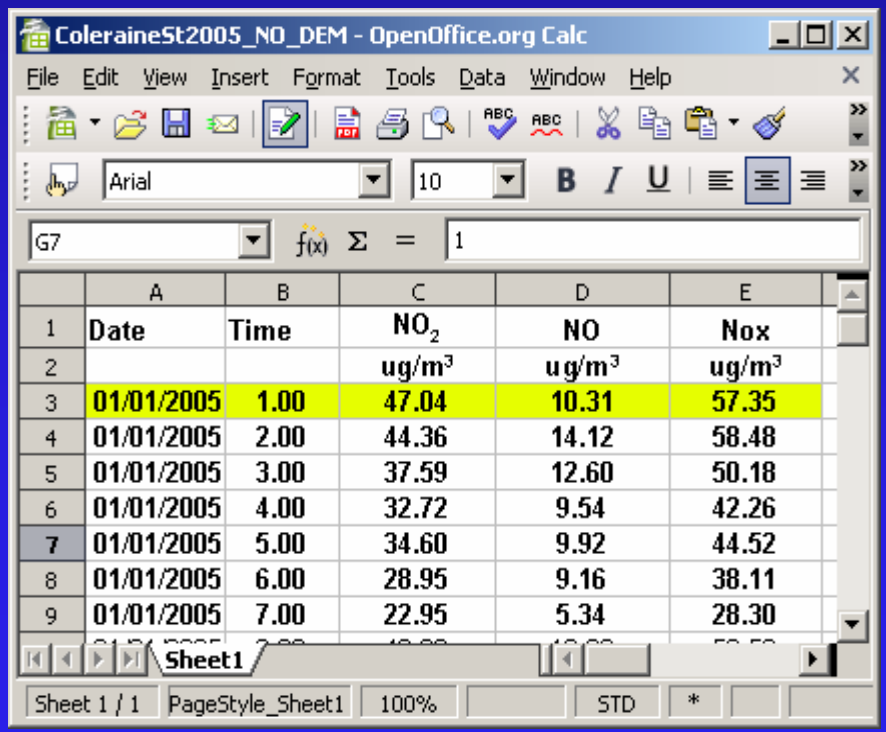
- Open Document Formats for Office Documents
- Document Content Stored in XML – easily parsed



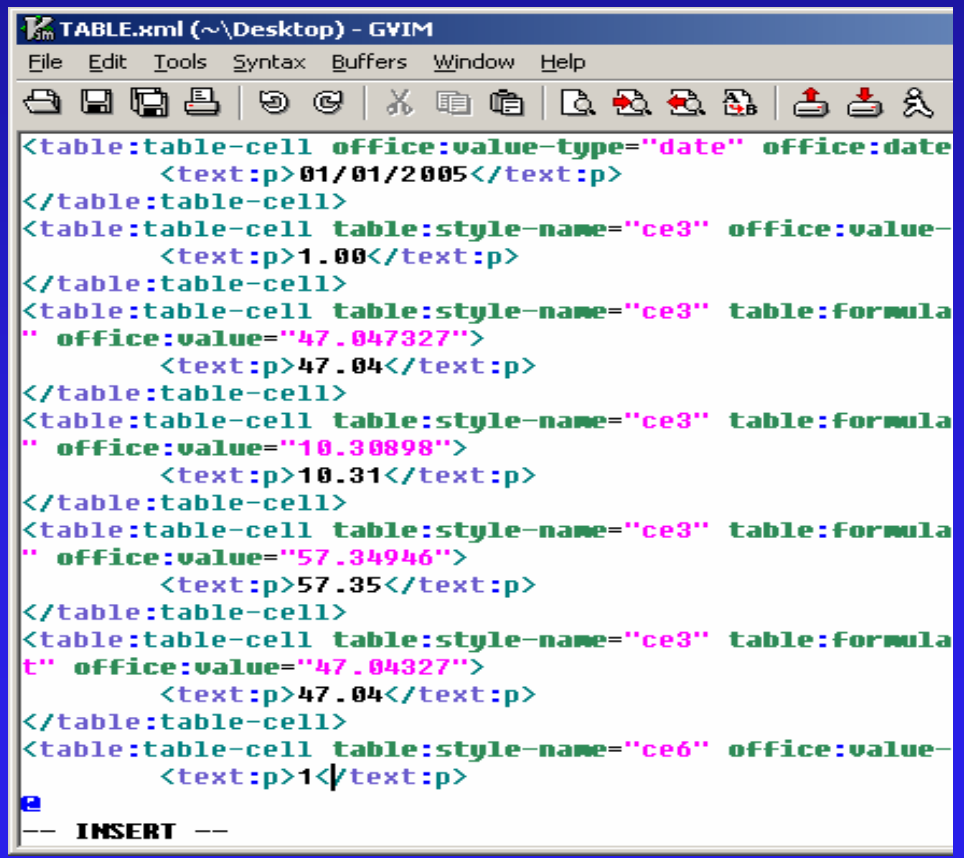


Open Documents Permit Sophisticated Parsing and Data QA/QC

- The ODS XML is very verbose for automated parsing
- More opportunities for better “data cleansing” (QA/QC)



	A	B	C	D	E
1	Date	Time	NO_2	NO	Nox
2			ug/m³	ug/m³	ug/m³
3	01/01/2005	1.00	47.04	10.31	57.35
4	01/01/2005	2.00	44.36	14.12	58.48
5	01/01/2005	3.00	37.59	12.60	50.18
6	01/01/2005	4.00	32.72	9.54	42.26
7	01/01/2005	5.00	34.60	9.92	44.52
8	01/01/2005	6.00	28.95	9.16	38.11
9	01/01/2005	7.00	22.95	5.34	28.30



Some Conclusions…..

Ensuring Data Interoperability mixes technical + non-technical approaches

- Offer support to choose best data management solution at project outset.
- Help to “train” researchers into good data management practices
- **Gain Researcher Trust**:
  - by showing how useful data sharing is to the scientific community
  - Explaining the security features of the system

OGC Services greatly simplify data reporting and data exchange

- Data is maintained in ONE place only
- Advanced query functionality available
- Open access interface to ANY software implementing OGC specifications
- On-the-fly data conversion + data mapping

Some Acknowledgements

Funding Position Code
EPA 2002-CC-FS4-MS4

Questions

or..
More Information …..

Peter Mooney
Email: peter.mooney@nuim.ie