

# Ontologies for e-Research

Dr Alun Preece  
Computing Science, University of Aberdeen



## An ontology is...

A term co-opted by computing science from philosophy.

Originally dealt with the nature and organisation of reality.

Now refers to an engineered artifact:

- a vocabulary denoting “things” in a particular reality
- a formalisation of the intended meaning of that vocabulary

The classic definition (after Gruber 1993):

- “An explicit specification of a (shared) conceptualisation”



## An ontology is useful for...

Communication, communication, communication!

Associating data with a (shared) vocabulary

- e.g. Gene Ontology used to “mark-up” biological datasets

Sharing data (and other artifacts)

- for data sharing, plays a role similar to (and better than?) an integration schema in databases
- (BUT this does not mean that a DB schema is an ontology!)
- a key issue is in aligning and linking multiple ontologies...

Human-computer interfacing

- information architecture: speaking the user’s language

Making software services interoperate

- machine-to-machine communication



## The Semantic Web

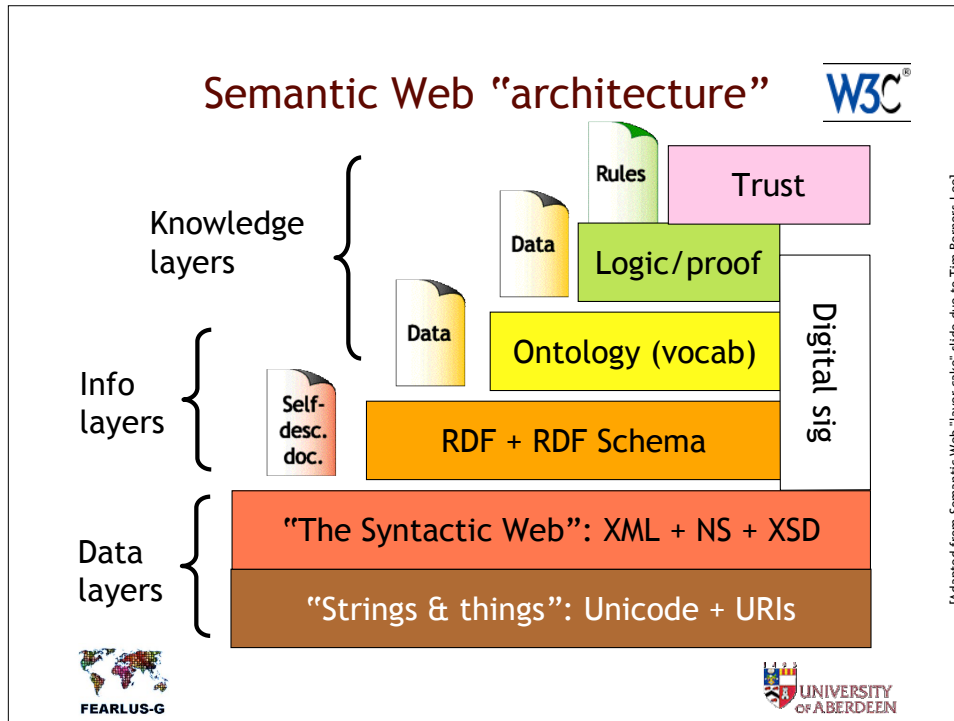


“An extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

“It is the idea of having data on the Web defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across various applications.”

Tim Berners-Lee, James Hendler and Ora Lassila  
*The Semantic Web*, Scientific American, May 2001





## e-Research, Grid, & Semantic Grid

e-Research is "the large scale research that is increasingly being carried out through distributed global collaborations enabled by the Internet."

[adapted from NeSC's e-Science definition]

The Grid is "an infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources."

[Foster & Kesselman]

The Semantic Grid is "an extension of the current Grid in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

[adapted from W3C's SW definition]



## Some things in the domain of e-Research

### Publications

- formal/reviewed
- "grey"
- associated artifacts

### People

- expert directories
- communities of practice

### Projects

- formal/funded
- working groups

### Software

- modelling & simulation
- number-crunching

### Experiment datasets

- formally curated
- raw/pre-processed
- *in vivo* / *in vitro* / *in silico*

### Scientific method

- experiment workflow
- knowledge roles:  
hypotheses, observations,  
predictions, ...
- discourse & argument

Ontologies!



## Metadata example: Dublin Core



The **Dublin Core Metadata Initiative** originated with the library community, intended to cover the properties of information artefacts in a library (including digital libraries).

The DC element set is a (weakly-specified) vocabulary defined within the XML namespace <http://purl.org/dc/elements/1.1/> (conventionally prefixed "dc:").

Examples:

title	date
creator	type
subject	format
description	language
publisher	<i>and much more...</i>

Most information resources in e-Research have these common properties, so the Dublin Core vocabulary has wide applicability.



## Dublin Core in RDF/XML

```

<?xml version = "1.0"?>
<rdf:RDF xmlns:rdf =
    "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description
    rdf:about="http://www.csd.abdn.ac.uk/~apreece/talks/
    OntologiesForERResearch">
    <dc:description>Slides from talk given to the CAVES ontology
    workshop at the Macaulay Institute</dc:description>
    <dc:creator
      rdf:resource="http://www.csd.abdn.ac.uk/~apreece"/>
    <dc:date>2005-06-21</dc:date>
  </rdf:Description>
</rdf:RDF>
  
```



## Metadata on people & projects: example

### AKT CS AKTive Space Take a tour through CS AKTive Space

About this page  research area/region  region/research area

**Research area**

- Theory of Computation
  - mathematical logic and formal languages
  - logics and meanings of programs
  - analysis of algorithms and problem compl
  - computation by abstract devices
  - general
- Mathematics of Computing
  - probability and statistics
  - discrete mathematics
  - numerical analysis
  - general
- Information Systems
  - information interfaces and presentation
  - information systems applications
  - information storage and retrieval
  - database management
  - general
- Computing Methodologies
  - document and text processing
  - simulation and modeling

**Radial:** 50 miles

**Map:** uk-political

**Researcher**

Top  5  10  20  unlimited

Order by  Grant total  RAE result

- KN Brown
- P Edwards
- AD Preece
- TJF Norman
- JRW Hunter
- PMD Gray
- DH Sleeman
- ED Reiter
- MW Freeston
- PJF Lucas
- GJL Kemp



## Metadata on people & projects: issues

### Far less standardised than publications

- no equivalent to Dublin Core, though FOAF (Friend-Of-A-Friend) is gaining ground
- several "portal schemas" in substantial use (including the AKT Portal Ontology, defined in OWL Full)
- little interoperability

### CAS is mainly populated by harvesting data

- sites don't provide it
- when they do, it isn't in the right format
- the 90/10 issue is key

### Named entity reconciliation is a big problem

- e.g. "Alun Preece" vs "A Preece" vs "A D Preece"

### Provenance & information quality (always)



## Managing experiment datasets

### Sticks & carrots:

- an increasing number of journals (e.g. in biology) require published datasets
- funding councils are becoming more concerned with reusability of results
- standard data formats are becoming more common, especially those based on XML
- datasets share some metadata characteristics with other published artefacts

### Issues:

- capturing context, for reuse of data
- cost of re-use compared to simply re-generating the data
- inevitably, provenance & information quality

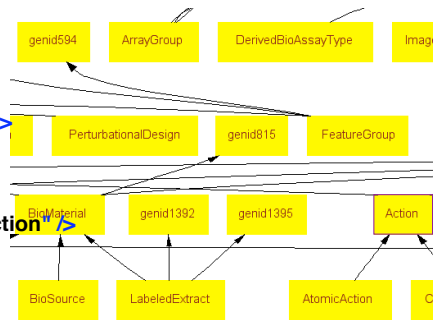


## Associating datasets with ontologies

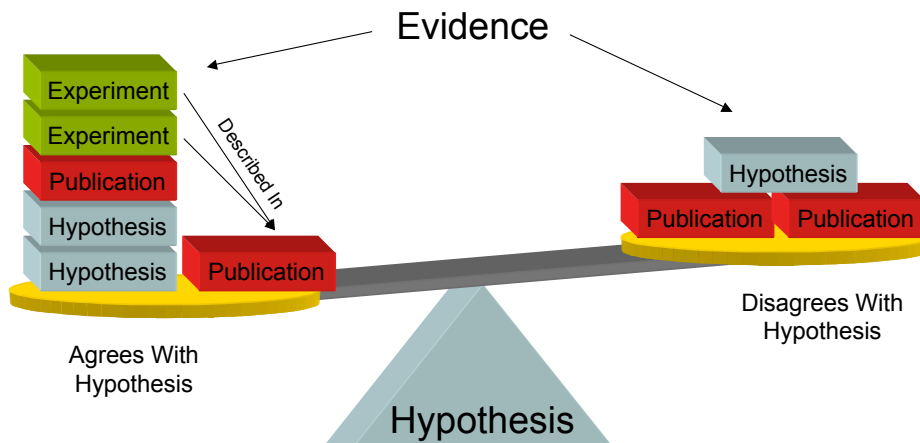
```

<BioSample
  identifier="S:Sample:MEXP:167278"
  name="CH131_1">
  <MaterialType_assn>
    <OntologyEntry
      category="MaterialType"
      value="whole_organism" />
  </MaterialType_assn>
  <Treatments_assnlist>
    <Treatment order="1"
      identifier="T:Sample:MEXP:167278">
    <Action_assn>
      <OntologyEntry
        category="Action"
        value="specified_biomaterial_action" />
      </Action_assn>
    </Treatment>
  </Treatments_assnlist>
</BioSample>
  
```

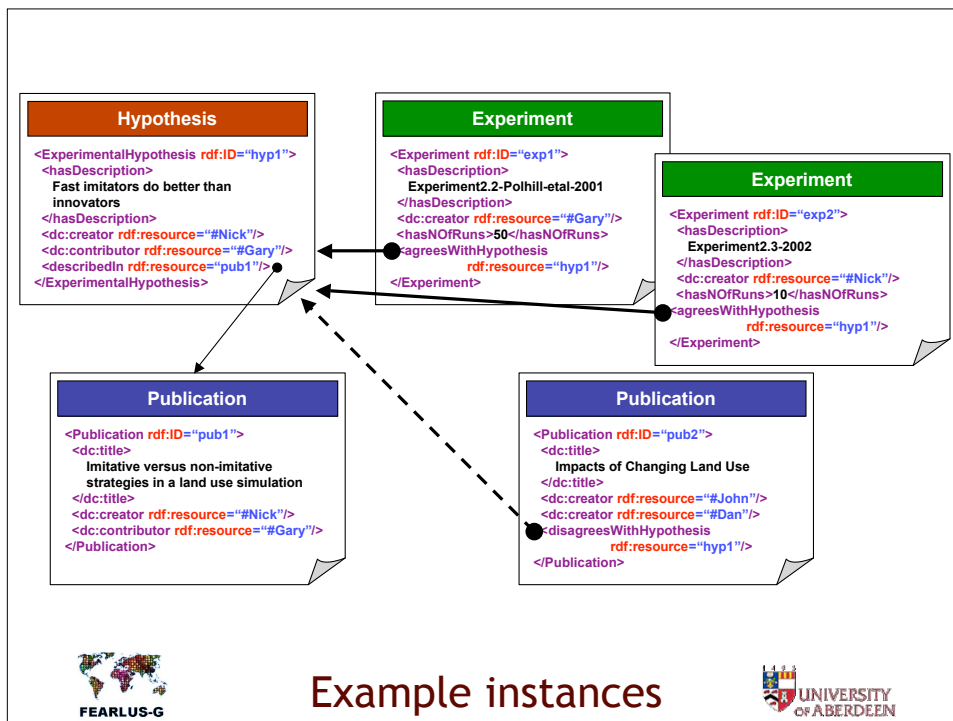
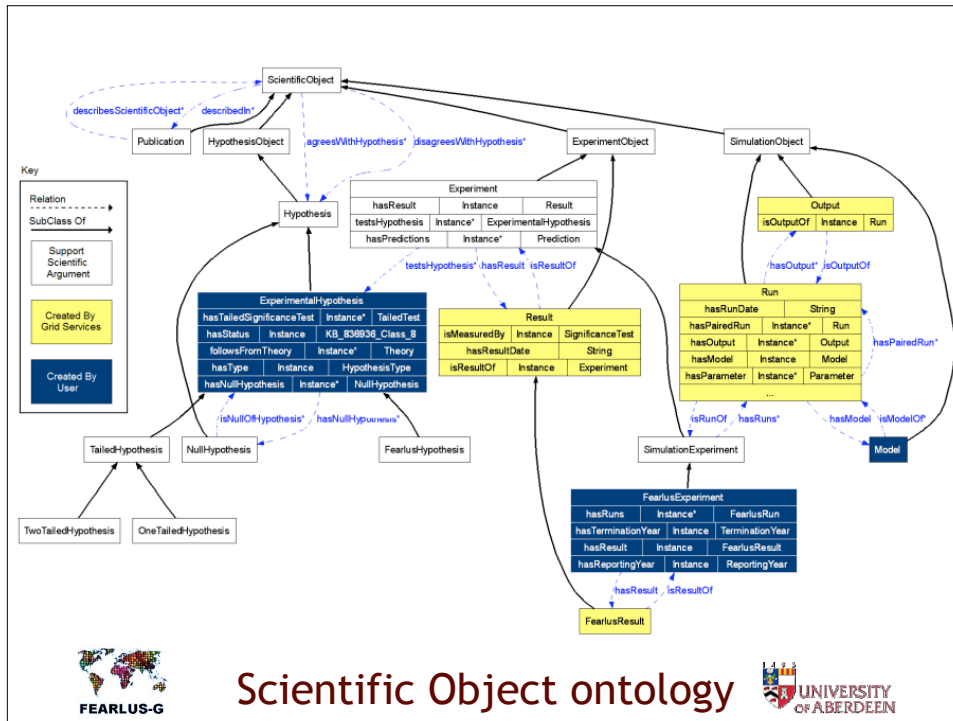
It's now common for datasets in biology to refer to ontologies.  
Example: MIAME / MAGE Ontology



## Objects in evidence-based science



The Semantic Grid / e-Research infrastructure





**Tools 1: Protégé**

**Tools 2: Longwell**

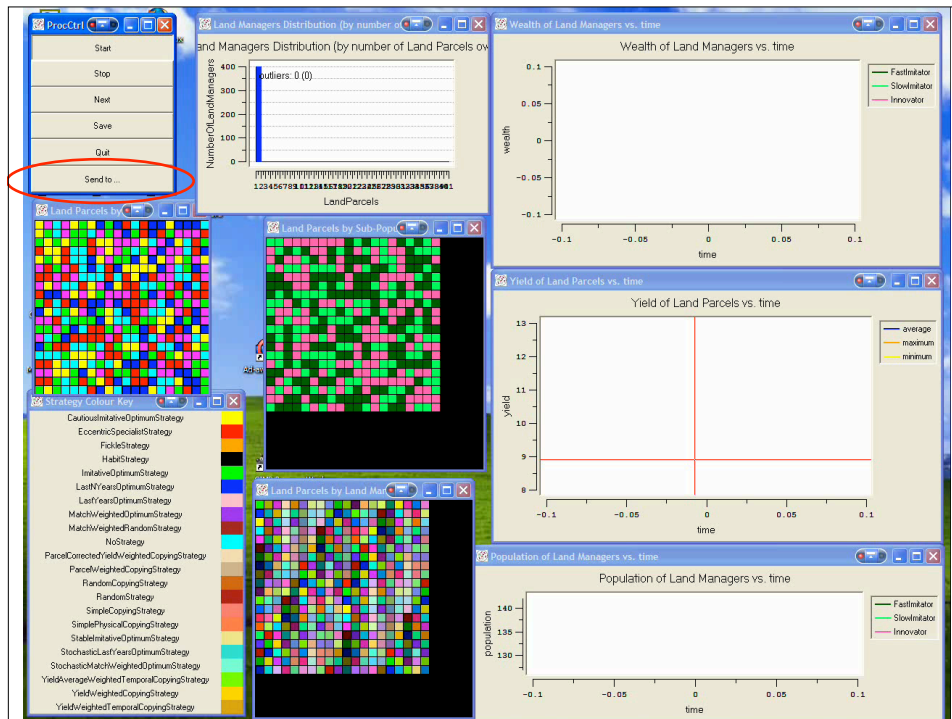
## Aims of the Fearlus-G project

To serve FEARLUS, an existing environmental modelling framework, to the scientific community


- allow very large-scale experiments to be run, analysed, and repeated

To promote collaboration by facilitating access to alternative models and comparison of results


To support training by providing a shared co-laboratory environment for experimentation



UNIVERSITY of ABERDEEN Computing Science



## Pilot Semantic Grid Service for Environmental Modelling






[Home](#)   [Upload](#)   [Search](#)   [Experiment](#)   [Log Out](#)


### Welcome to the FEARLUS-G Web Client

Models Parameters uploaded from: ed

Model Parameters File Name	Date	Time	Description	Status	Options
Fi-v-Si-v-In.model	2004-07-29	16:23:57	Innovator vs. Imitators	Completed	<a href="#">Run</a> <a href="#">View</a> <a href="#">Delete</a> <a href="#">Publish</a> <a href="#">Clone</a> <a href="#">Send</a>
P0-E16u-BET8-H8RvH12R.model	2004-08-20	15:12:57	Innovators vs Imitators	Running...	<a href="#">Run</a> <a href="#">View</a> <a href="#">Delete</a> <a href="#">Publish</a> <a href="#">Clone</a> <a href="#">Send</a> <ul style="list-style-type: none"> <li>▶   Simulation instance 0 running...</li> <li>▶   Simulation instance 1 running...</li> <li>▶   Simulation instance 2 running...</li> <li>▶   Simulation instance 3 running...</li> <li>▶   Simulation instance 4 running...</li> <li>▶   Simulation instance 5 running...</li> <li>▶   Simulation instance 6 running...</li> <li>▶   Simulation instance 7 running...</li> <li>▶   Simulation instance 8 running...</li> <li>▶   Simulation instance 9 running...</li> <li>▶   Simulation instance 10 running...</li> <li>▶   Simulation instance 11 running...</li> <li>▶   Simulation instance 12 running...</li> <li>▶   Simulation instance 13 running...</li> <li>▶   Simulation instance 14 running...</li> </ul>


FEARLUS-G  
My Workspace




#### My Projects

 [Imitative Versus Non-Imitative Strategies](#)


[+ New Project](#)

#### My Experiments

 Experiments

<input checked="" type="checkbox"/> SI-v-HRYI-c-II-env1-set2.2	<input checked="" type="checkbox"/> SI-v-II-c-HYI-env1-set2.2
<input checked="" type="checkbox"/> SI-v-II-c-HRYI-env1-set2.2	<input checked="" type="checkbox"/> II-v-HYI-c-SI-env1-set2.2
<input checked="" type="checkbox"/> SI-v-HYI-c-II-env1-set2.2	<input checked="" type="checkbox"/> II-v-HRYI-c-HYI-env1-set2.2
<input checked="" type="checkbox"/> SI-v-HRYI-c-HYI-env1-set2.2	<input checked="" type="checkbox"/> II-v-HRYI-c-SI-env1-set2.2
<input checked="" type="checkbox"/> HVI-v-HRYI-c-SI-env1-set2.2	<input checked="" type="checkbox"/> HVI-v-HRYI-c-II-env1-set2.2
<input checked="" type="checkbox"/> II-v-HYI-c-HRYI-env1-set2.2	<input checked="" type="checkbox"/> SI-v-HYI-c-HRYI-env1-set2.2


Experiment Sets

 [Experiment2.2-Pohillletal-2001](#)



[+ New Experiment](#)   [+ New Experiment Set](#)

#### My Simulation Parameters









**Models**

 [modelo-G-6unix](#)

**Environments**

 [Environment1](#)    [Environment2](#)

**Subpopulations**

 <a href="#">OD</a>	 <a href="#">OS</a>	 <a href="#">HYI</a>
 <a href="#">RS</a>	 <a href="#">II</a>	 <a href="#">SI</a>
 <a href="#">LS</a>	 <a href="#">HRYI</a>	

[+ New Model](#)

#### Options

[My Workspace](#)  
[My Repository](#)

[Log Out](#)

#### FEARLUS-G Options

[Upload a model form file](#)  
[Search a model in MyWorkspace](#)

#### Experiments

[Type 1 Experiment](#)

**Experiment Set**

**Experiment2.2-Polhill-etal-2001**

Label: Experiment2.2-Polhill-etal-2001

Description: Experiment2.2-Polhill-etal-2001

Comment:

Number of Runs: 60

Significance Level: 0.0010

Single Subject: true

Hypotheses:

Experiments:

- II-v-HRYI-c-SI-env1-set2.2
- SI-v-HYI-c-II-env1-set2.2
- II-v-HYI-c-HRYI-env1-set2.2
- HVI-v-HRYI-c-SI-env1-set2.2
- SI-v-HRYI-c-II-env1-set2.2
- SI-v-HYI-c-HRYI-env1-set2.2
- II-v-HRYI-c-HVI-env1-set2.2
- SI-v-HRYI-c-HVI-env1-set2.2
- II-v-HYI-c-SI-env1-set2.2
- SI-v-II-c-HYI-env1-set2.2
- SI-v-II-c-HRYI-env1-set2.2
- HVI-v-HRYI-c-II-env1-set2.2



Uses Model: modelo-6-6unix

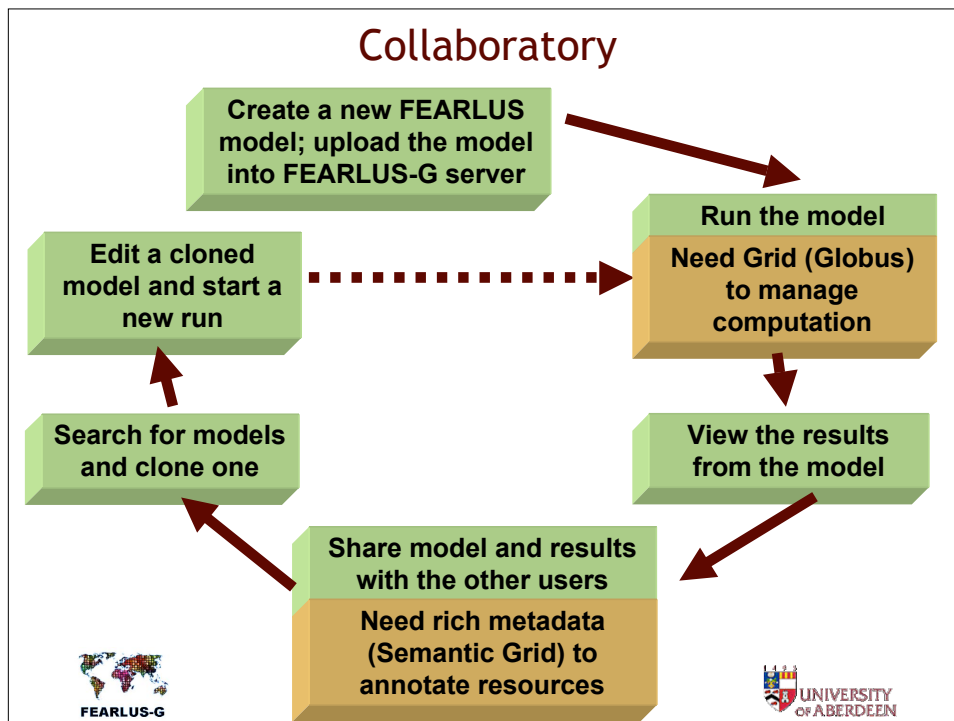
Termination Year: year201Termination

Subjects: HRYI, HYI, SI, II, Environment1

Hypothesis Type: fearlusType2Hypothesis

Tailed Hypothesis Class: OneTailedHypothesis



## Some challenges

### The “annotation bottleneck”

- getting researchers to mark-up their resources
- NLP can help to some extent (e.g. Sheffield’s Armadillo)
- also ontology search & matching

### Semantic disambiguation

- recognising resource X and resource Y are the same thing
- the need for URI schemes & use of OWL: sameAs, func props...

### Scaling-up

- managing huge numbers of RDF statements (e.g. 3store)
- inference!

### “Semantic cracks” - fill them or just paper-over?

- XML/RDF/OWL/rules layering is broken
- but can we make “The Pragmatic Web” work anyway?



## Links & credits

### See also

- [www.csd.abdn.ac.uk/research/fearg](http://www.csd.abdn.ac.uk/research/fearg)
- [www.aktors.org](http://www.aktors.org)

### Fearlus-G people

- Pete Edwards (Aberdeen)
- Edoardo Pignotti (Aberdeen)
- Nick Gotts (Macaulay Institute)
- Gary Polhill (Macaulay Institute)



... any questions?

