PEG-BOARD
An e-Science Case Study

Gregory J. L. Tourte Emma Tonkin Paul Valdes

g.j.l.tourte@bristol.ac.uk
e.tonkin@ukoln.ac.uk
p.j.valdes@bristol.ac.uk

ELPUB 2010 – Helsinki, Finland — 2010–06–17

Introduction

- Data management and curation of Palæoclimate data produced by BRIDGE research group
- In collaboration with University of Bath (UKOLN), University of Leeds (Earth Science) & University of Southampton (Archæology)
- Data centered around climate model runs and outputs from HPC facilities (institutional and national)
- This includes model's binaries, initial conditions, forcings, model outputs, graphical analysis
- Very diverse user base, not only palæoclimate researchers, but geologists, biologists, archæologists, media,. . .

About BRIDGE

- BRIDGE: Bristol Research Initiative for the Dynamic Global Environment.
- Research group with the School of Geographical Sciences at the University of Bristol, UK.
- focused on Palæoclimate research to
  - understanding the Earth system
  - validating current climate models
  - improving models by incorporating earth system variables
  - providing data to other disciplines
- What is involved
  - Climate model (GCMs)
  - HPC infractructure (national & intitutional facilities)
  - Data management (produce terabytes of data per day)

Making data available

- diverse audience and user base
  - different backgrounds
  - different use for the data
  - different skillset
  - different requirements
- partial solution: national data centres funded by NERC
  - have existed for quite some time
  - but fairly restrictive
  - tend to be targeted at climate researchers
  - does not provide analysis tools, just access to stored data in raw format

The PEG–BOARD project

- extension, rationalisation and standardisation of toolset created by Paul Valdes
- started as ad-hoc series of script to help personal workflow
- grew to ease access to data by collaborators
- became main tool for processing data within the research group, used by postgraduades and RA to process output.
- however, technology evolutions means more data generated.
- ad-hoc solution not sufficient.
- need for management, preservation and curation

Aims

- Assist work of modellers
- facilitate re-use
- enable discovery
- provide data retention policies and guidelines
- ensure data retention & curation policies reflects research and data life cycle.

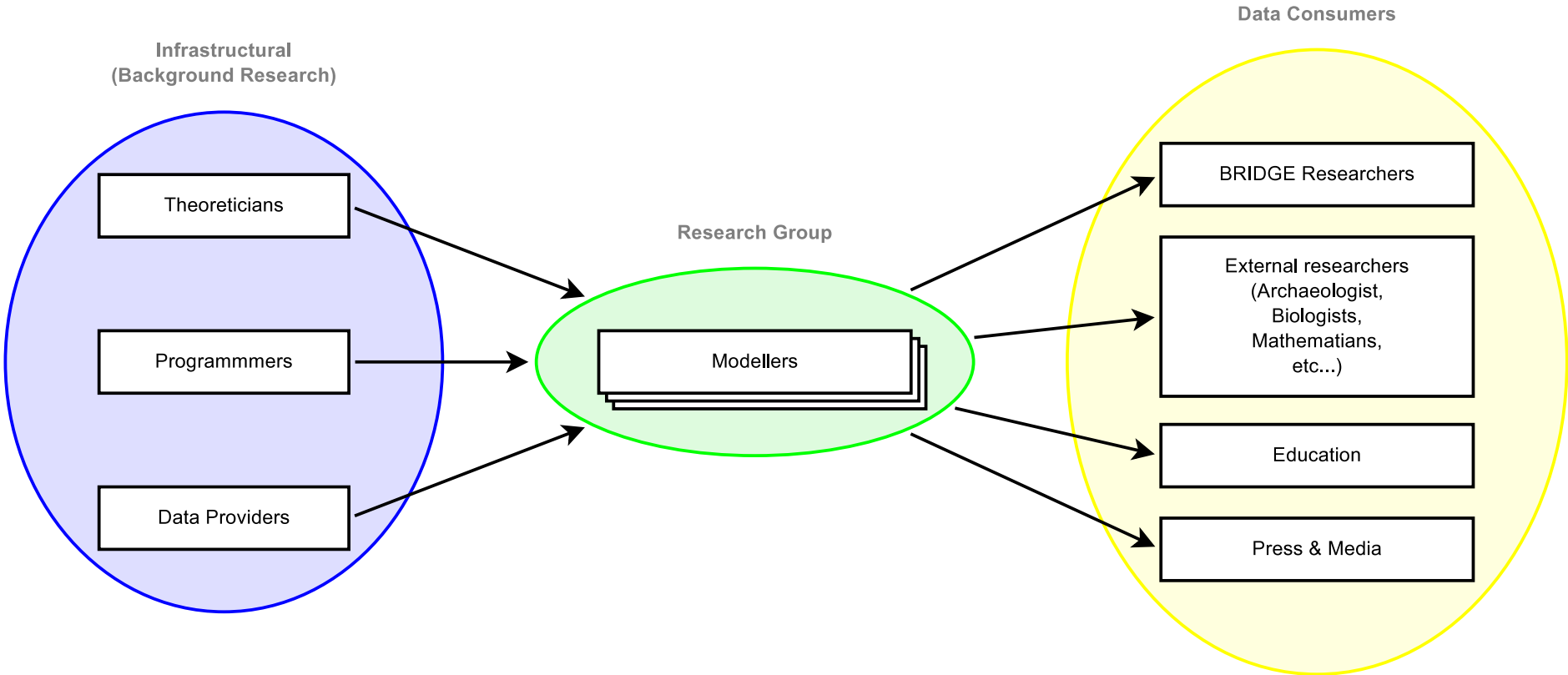
Methodology

- Pre-existing infrastructure
- Knowledge of existing code
- Day-to-day maintenance and evolution of user needs
- User interviews (informal, using users' support requests)
- User survey (still on-going) using modified DAF to include external users

Stakeholder Analysis

- Worked jointly with HPC group with University wide storage project
- Devided users into roles (Swann, 2008)

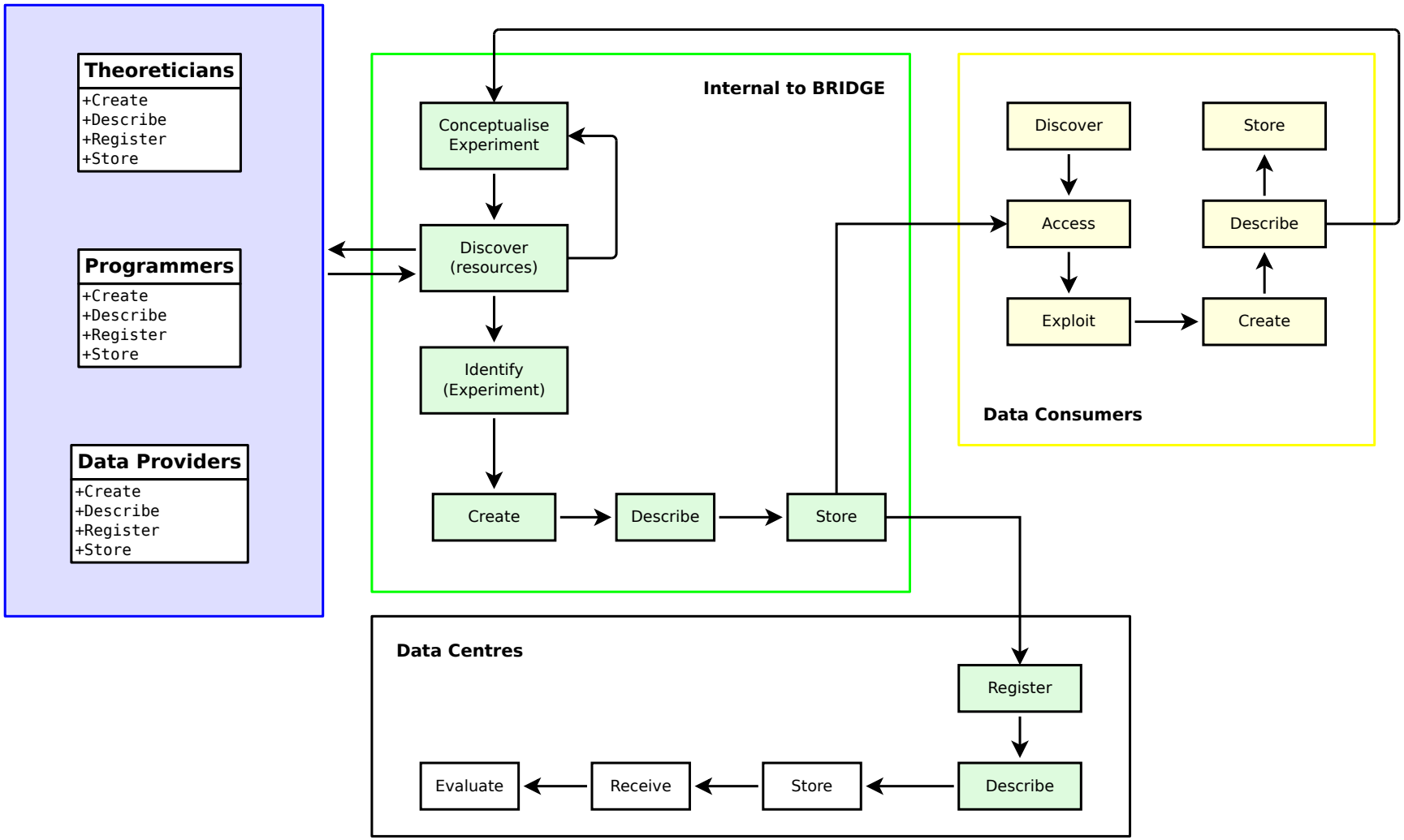
Stakeholder Analysis – Part 2



Task Based Analysis

- Role based model did not fit with reality
- Users have more than one role
- Interactions between roles and users can be complex
- Tasks analysis approach more appropriate (Based on Treloar, ANDS, 2009)

Task Based Analysis – Part 2



Case Study: Archæology

- non-computer savvy
- not familiar with raw data output (netCDF)
- different ontology to describe things compared to modellers
- more data available than is required
- ambiguous variable desciptions
- different units used
- need for pre-processed output mainly in the for of plots rather then raw data.
- need for a clean & easy way to identify which experiment is needed without the modeler to be required

Upgrading BRIDGE

- huge task ahead
- focus first on management/requirement analysis
- preservation; accessibility; metadata extraction
- automated metadata extraction to process past experiments
- accessibility & visualisation

The End...

Any Questions?