PURE Data Challenge

Bobby Stuijfzand

Data Science Specialist

Jean Golding Institute

Jean Golding Institute

- Bringing together the University’s strengths in data-intensive research.
  - Fostering a cohesive community
  - Showcasing current work internally and externally
  - Offering help and advice
  - Providing new learning opportunities for staff and students

Famous for Three Minutes

- Researchers from different schools talk about their data-intensive research in three minutes
- Networking afterwards, i.e. free lunch and chats

Data is Beautiful

Jean Golding Institute
for data-intensive research
Making data work for everyone
Nathan Chrismas
PhD. student, Geographical Sciences
(Glaciology), Sciences
Macrocyanobacteria
WINNER

Ask JGI

- Free advice and support on data science and signposting to experts and facilities within the University
  - Statistics
  - Computing
  - Data management
- Email support, 1-1 meetings, grant support
- ask-jgi@bristol.ac.uk

The PURE Data Challenge

- **Using data science to identify and analyse interdisciplinary research at Bristol**
  - How would we best visualise the data in PURE and how might we best use it to answer questions like which disciplines/schools are internally best connected in terms of research?
  - Which individuals/groups bridge the divides between schools and faculties?
  - If we were to take the PURE data and attempt to cluster the individuals into groups would these groups resemble the existing school / faculty structures? If not, what would they represent and would they be useful representations?
  - Could the data identify gaps in the interdisciplinary landscape that might be usefully filled to prepare Bristol for cross-disciplinary calls?
  - What is the best way to measure interdisciplinary research?
  - Can the team detect potential collaborations and/or overlapping research areas that have not been identified in the past (via research groups, institutes, centres, etc)?
  - Can we develop an advanced form of visualisation that complements PURE and Scival?

The Data

- Three csv files containing all PURE research outputs up until last REF submission (2014)
  - Table with outputs
  - Table with staff information
  - Table with staff and output identifiers
- Provided by PURE technical team and JGI’s data scientist (yours truly)
- Made available through fluff link

  (make sure to provide email address!)

Outputs.csv

- **publication ID**
- title
- type of publication code
- type of publication
- publication day
- publication month
- publication year
- abstract
- keywords (all keywords for a publication are included in the same cell, comma separated)

Staff.csv

- **person id**
- published name
- forename
- surname
- organisational code
- job title

Authors.csv

- **person ID**
- **publication ID**
- published name

Data characteristics

- About 34000 research outputs
- About 29000 of these with author information
- BUT: Messy data
  - Some duplicates (about 200)
  - Direct extract from PURE, so be mindful of escape characters etc.
  - Keywords!

Competition Rules

- **Deadline of submission: 19th May 2017**
- **Point of submission:** ask-jgi@bristol.ac.uk
- Prizes: £1000 for the winner, 2x £250 for runners up
- Sign up at any point during competition
- Up to 5 people can be part of a team.
- Winning team announcement: 6th June 2017
- Winners' showcase: 12th September 2017 (tbc)
- The Jean Golding Institute have the rights to publicly disseminate any entries.