A case study in the rewards of long term data sharing :

26 years of the MRC Psycholinguistic Database

Michael Wilson
STFC Rutherford Appleton Laboratory

Science & Technology Facilities Council
e-Science

The message

Publishing data long term
with an accompanying journal publication
can produce a large number of citations
which will be good for your career
in the Research Excellence Framework
so **its worth publishing data**

Its not worth archiving data

Research councils have established data sharing policies which require data to be made available by researchers and curated for future use by others.

It is argued that nobody will use this data and that it's too much effort to complete the metadata required by archives to index it.

But it is ….

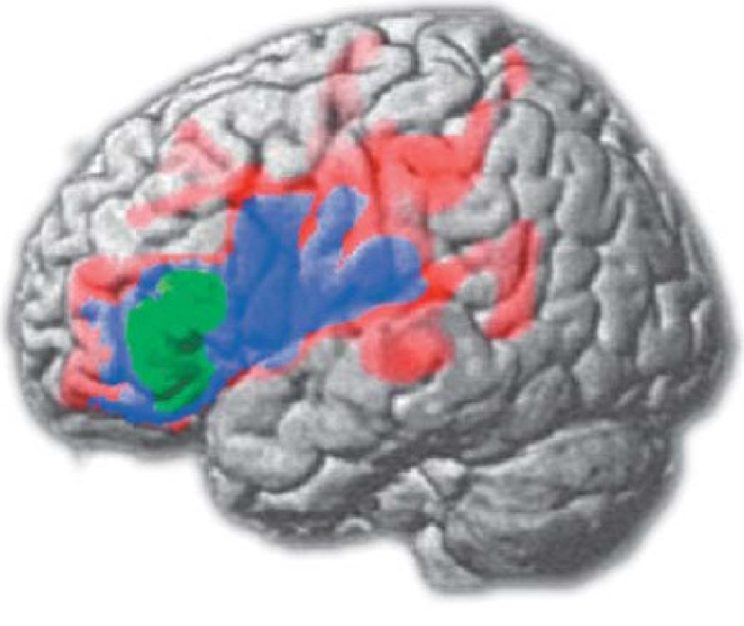
MRC Psycholinguistic Database

- Single 11 Mbyte sequential UNIX file
- 150837 words
- 9 data sources
- information about 26 different linguistic properties
    - Number of letters, phonemes, syllables
    - Word frequency
    - Imagery, age of acquisition, etc…

Data Usage

To select stimulus words for experiments in psychology, neurology, linguistics

Experiments measure decisions, response time, brain activity



Technological evolution.

1981 – dataset made available as a postal access service to ULCC.

1988 dataset released through ftp access from the Oxford Text Archive so that users could operate it locally.

1994 the dataset was made available through web interfaces at RAL and the University of Western Australia for users to use remotely

Promotion and Reference

Access has been free, and there has been no significant promotion or advertising of the dataset.

Users of the dataset have been asked to cite one of two papers in their own publications (Coltheart, 1981; Wilson 1988) when they use the dataset

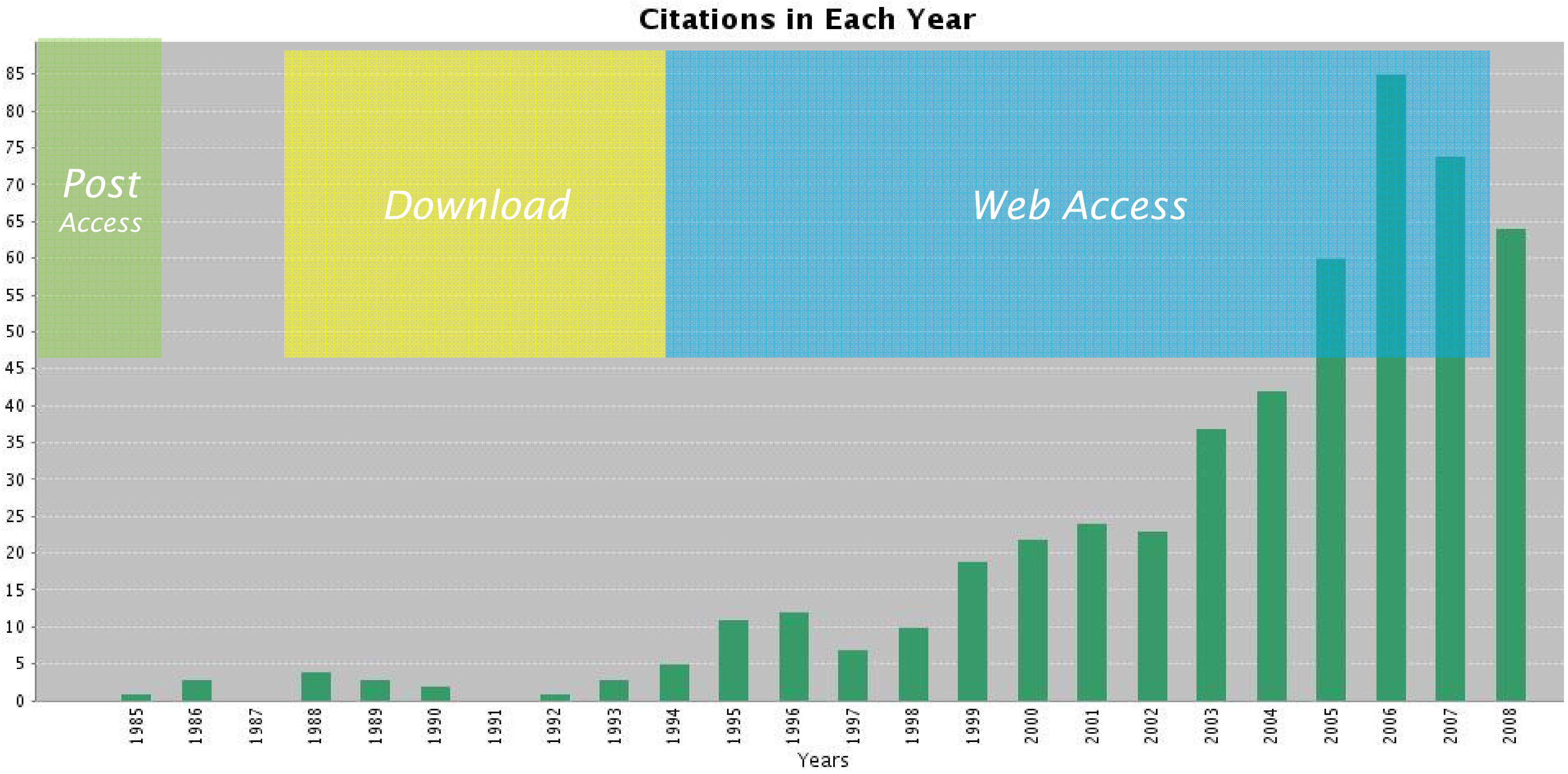
Sustainable curation

- *Data hosted on a supported web server*
- *About 1 week effort per year user and application support*

Impact: Usage and Citation

- *Usage statistics of free services don’t prove the seriousness of the users or the impact of their outputs*
    - *About 8000 queries in 2007*
- *Replacement of the UK RAE with a Research Excellence Framework (REF) will make greater use of citation indexes*
    - *About 500 citations in 26 years*

Citation history (1985 – 2008)-



Science & Technology Facilities Council
e-Science

Conclusion

Since

- *the data has not changed*
- *the datasets have not been recently advertised*

the explanation for this recent uptake is that

- *the culture of re-using data,*
- *the acceptance of web access to data,*
- *and the willingness to cite data sets*

have all contributed to the increase in citations.

Therefore, **those who archive their data from recent research should expect similar returns to benefit their REF evaluations.**

Recent Citation of Dataset Publication

At least **7 of the top 20** *UK-based Authors of High-Impact Papers, 2003-07 ranked by citations to high-impact papers published data sets:*

- *Pfam protein families database (> 1000 citations in 3 years)*
- *Sloan Digital Sky Survey,*
- *2dF Galaxy Redshift Survey*

References

COLTHEART M (1981) THE MRC PSYCHOLINGUISTIC DATABASE, QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY SECTION A-HUMAN EXPERIMENTAL PSYCHOLOGY Volume: 33 Issue: NOV Pages: 497-505.

WILSON M (1988) MRC PSYCHOLINGUISTIC DATABASE - MACHINE-USABLE DICTIONARY, VERSION 2.00, BEHAVIOR RESEARCH METHODS INSTRUMENTS & COMPUTERS Volume: 20 Issue: 1 Pages: 6-10

MRC Psycholinguistic Database – http://www.psych.rl.ac.uk

The U.K.'s Citation Elite, 2003-07 - ScienceWatch.com