Speech analysis and Archive research

Meta Forum 2012, Brussels

George Wright
Head of Internet Research & Future Services
BBC R&D

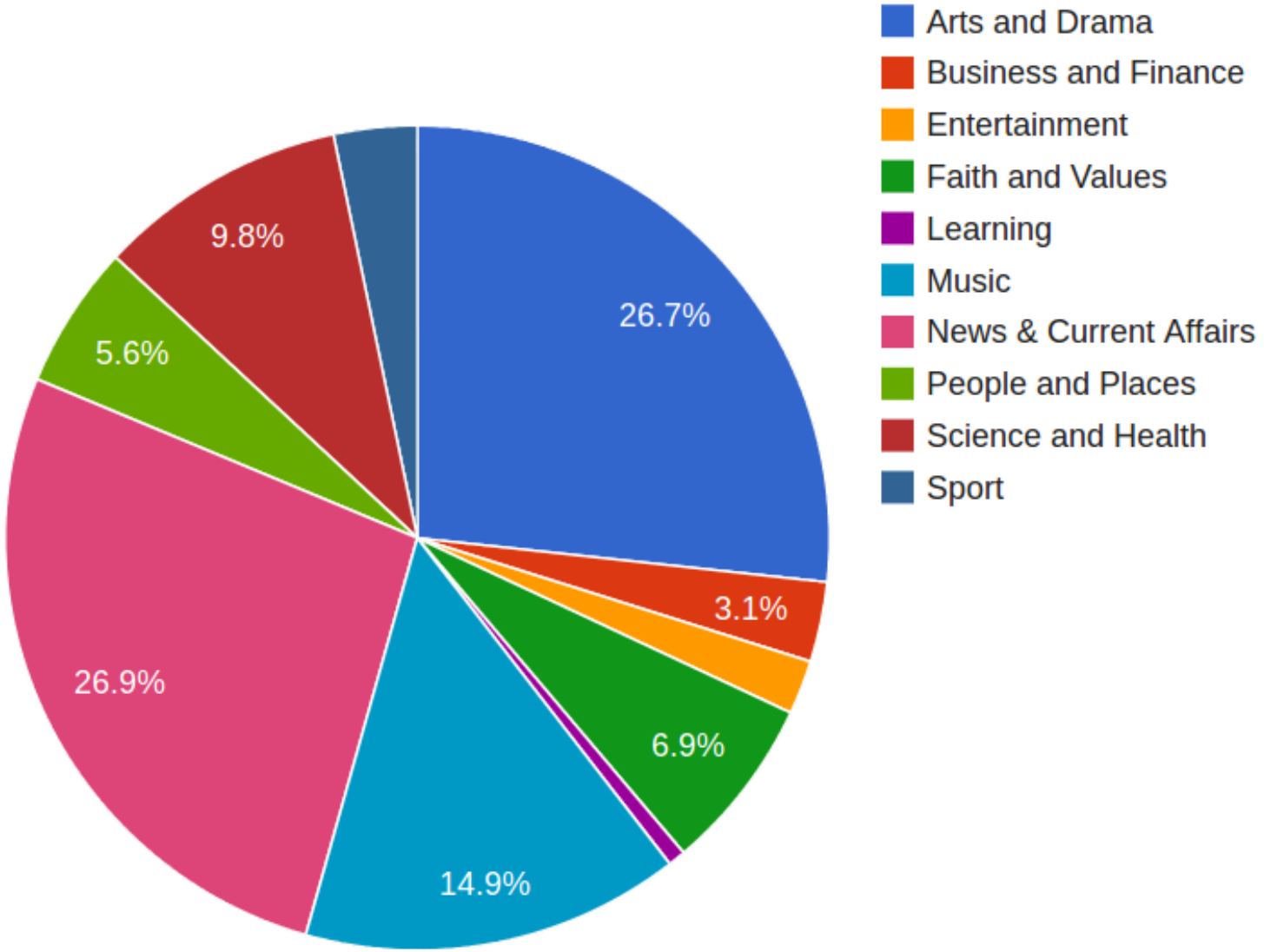
@georgie

BBC World Service archive

Around 70,000 radio programmes

- Covering +60 years
- ~ 3 years of continuous audio
- Sparse metadata
- ~ 500 TB of content

The Content by genre



The Content (Corpus)

All the BBC World Service's English language output

Plays, features, documentaries going back 60 years

Uncompressed WAVs and high quality MP3s

Rights cleared for worldwide use

Why we have to use language tools

- Impossible to re-categorise
- Some of this content very specialised domains
- Need to identify links within programmes, and to the web
- Ongoing growing archive
- Constant pressure to spend more money on programmes, less on programme 'management'

Why it's interesting and useful

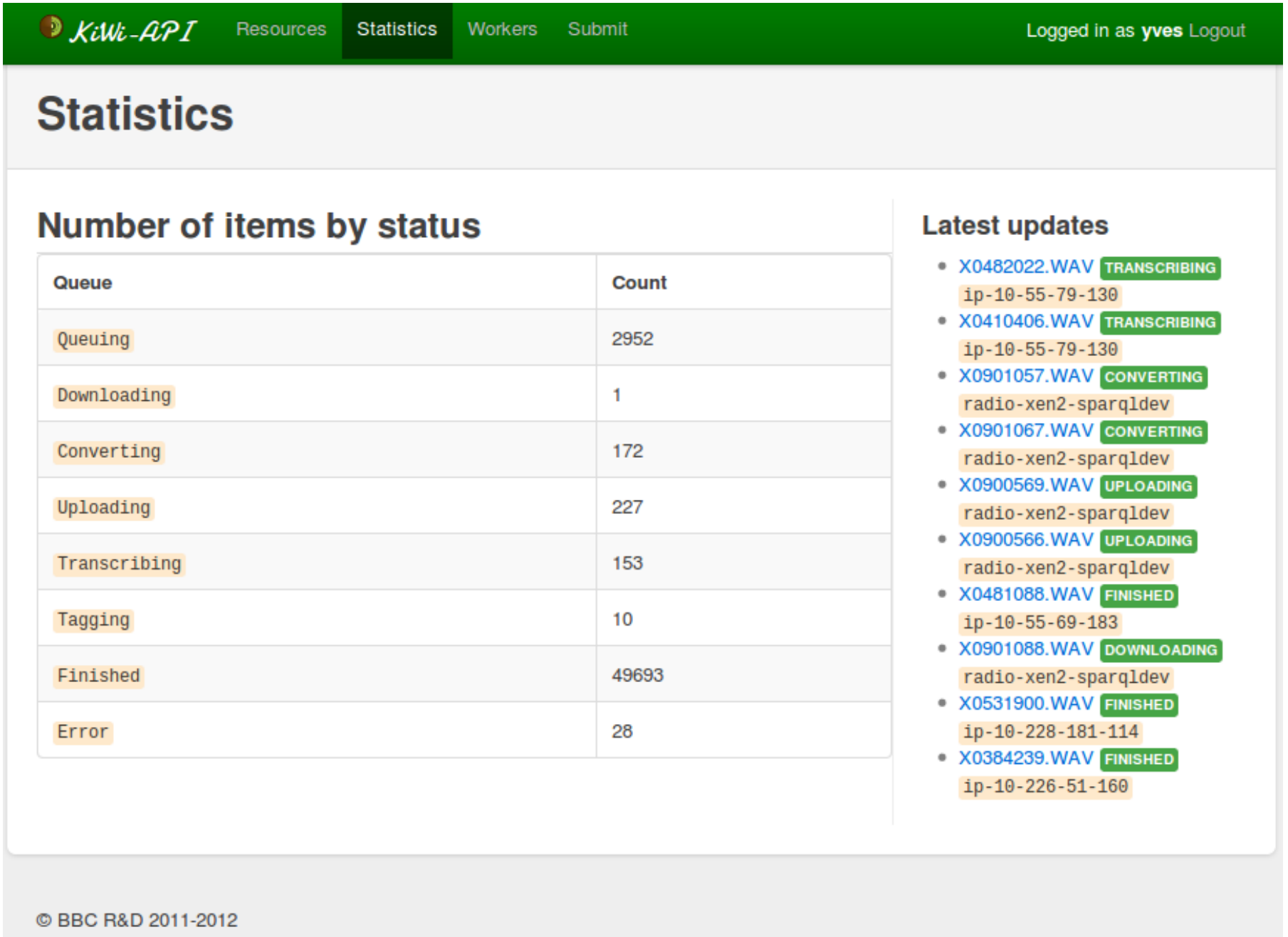
- Experimenting with a massive dataset
- Researching complex GUIs, worldwide content, interesting journeys
- Useful for other BBC content areas
- Useful for other domains

Why this is hard

- Lack of publicly available acoustic and language models for speech recognition.
- Most available acoustic models and language models (often under no clear licensing because derived from commercial datasets) are trained on US English, with US speakers.
- Real need for a good, open-source, British English acoustic model
- Need for good open language models for a variety of domains.
- World Service English unique!

Processing the World Service archive

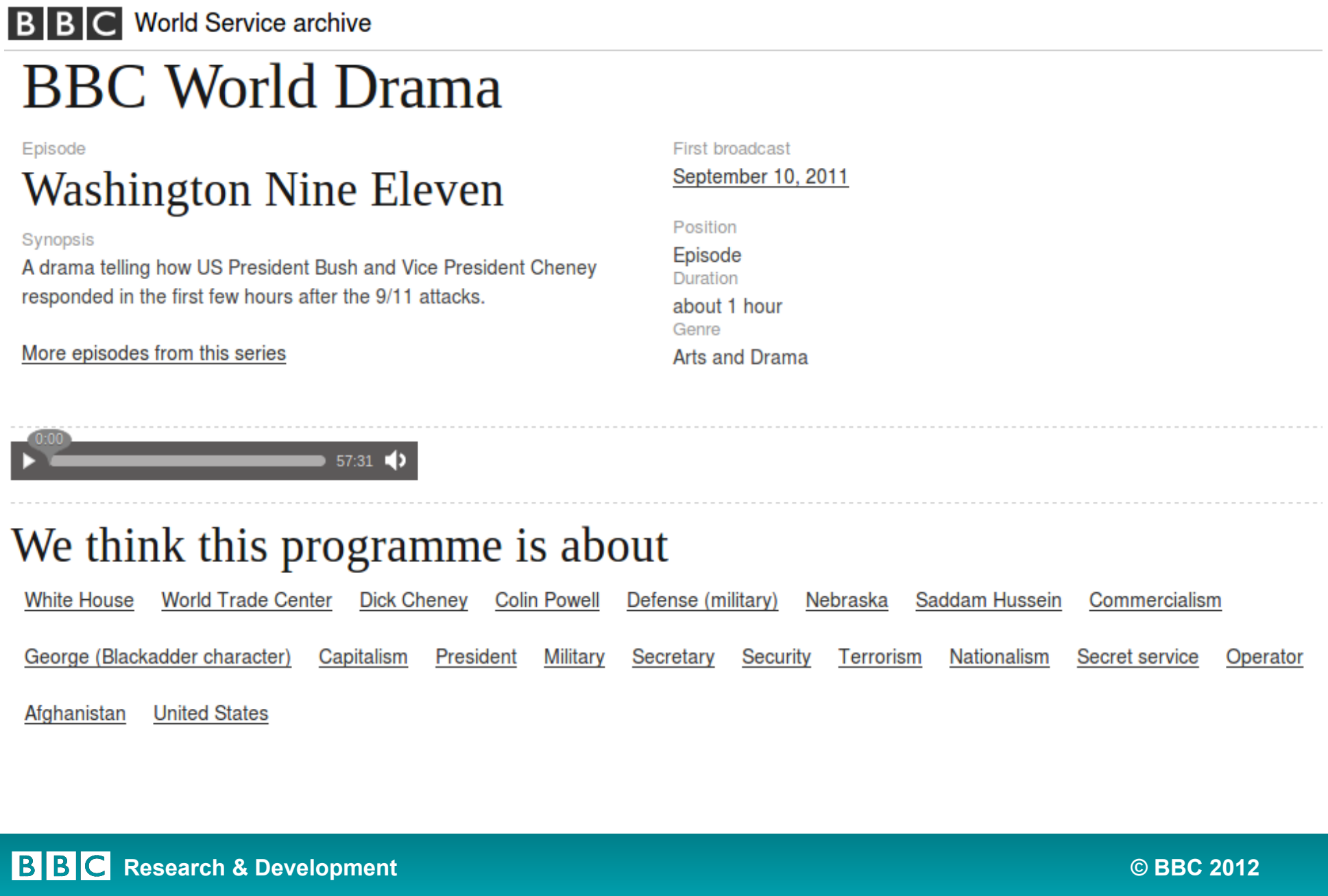
- Workflow steps isolated in individual workers
- Computation-intensive workers on Amazon Web Services
- All managed by Message Queues and an API centralising the data
- Only bottleneck: upload speed to Amazon's servers


KiWi-API
Resources
Statistics
Workers
Submit
Logged in as yves Logout
Statistics
Number of items by status
Queue
Count
Queuing
2952
Downloading
1
Converting
172
Uploading
227
Transcribing
153
Tagging
10
Finished
49693
Error
28
Latest updates
X0482022.WAV TRANSCRIBING ip-10-55-79-130
X0410406.WAV TRANSCRIBING ip-10-55-79-130
X0901057.WAV CONVERTING radio-xen2-sparqldev
X0901067.WAV CONVERTING radio-xen2-sparqldev
X0900569.WAV UPLOADING radio-xen2-sparqldev
X0900566.WAV UPLOADING radio-xen2-sparqldev
X0481088.WAV FINISHED ip-10-55-69-183
X0901088.WAV DOWNLOADING radio-xen2-sparqldev
X0531900.WAV FINISHED ip-10-228-181-114
X0384239.WAV FINISHED ip-10-226-51-160
© BBC R&D 2011-2012

Speech recognition tools

- Using CMU Sphinx
- HUB4 acoustic model
- Language model derived from the Gigaword corpus
- Word Error Rate of about 47%

In action



In action

BBC World Service archive

Discovery

Episode

Smallpox

Synopsis

Should we deliberately make a living organism extinct? What if it's the Smallpox virus, that killed around 300 million people in the 20th Century. The cases for and against destroying the remaining stocks, from Edward Hammond & Raymond Weinstein

More episodes from this series

First broadcast

May 16, 2011

Position

Episode

Duration

17 minutes

Genre

Science and Health

We think this programme is about

Smallpox Public health Infectious disease Researcher Manufacturing Laboratory Weapon Genomics Disease Clinic Virus

Vaccine Immune system Organ (anatomy) United States Infection Donald Rumsfeld Professor Futurism Health

In action

BBC World Service archive

Portraits Of Our Time

Episode

Ayatollah Khomeini

Synopsis

A profile of Iran's religious leader, the Ayatollah Khomeini, who toppled the Shah of Iran in 1979 to create an Islamic republic. Presenter and compiler: Justin Phillips.

More episodes from this series

First broadcast

October 01, 1983

Position

Episode

Duration

16 minutes

Genre

People and Places

We think this programme is about

Revolution Ayatollah Margaret Thatcher Brother (2000 film) Portrait Islam Rome Newspaper Smile (The Beach Boys album)

Ambassador Tehran Iran Land reform Remote control Arrest Driving Mozambique Lifestyle (sociology) Television Prophet

Peer reviewed outputs

- Two papers at World Wide Web conference, April 2012
- One paper at Extended Semantic Web Conference, May 2012
- Two posters and one paper at IBC, 2012

Critical issues

- Currently around 40% accuracy
- Needs to get better at understanding complex vocab, unusual language
- Reaching limit of existing state of the art
- Embedding best-of-breed academics with team
- Plan for public trial
- Speech rec needs more help!

Thank you

- george.wright@bbc.co.uk
- http://bbc.in/ws_archive
- http://bbc.co.uk/rd/

Thank you

- george.wright@bbc.co.uk
- http://bbc.co.uk/rd/