

**PRESS
ASSOCIATION**

Elegant, frictionless
tools for metadata
management

Helen Lippell and Liz Perreau
ISKO UK 9 July 2013



Agenda

- PA
- Business buy-in
- Project drivers
- Technical background
- Business engagement
- What worked well
- What didn't work well
- Data considerations

PA

- Founded 1868
- Content
- Data and listings
- Services
- Meteo Group
- Data provider to 2012 Olympics
- Digital products and semantics

Images removed for copyright reason

Business buy-in

- Senior management
 - Cost savings
 - Product innovation
- Support active leadership in media semantics
 - Dynamic publishing
 - SNaP Ontology
 - Relationships with research and commercial partners

Project drivers

- Help journalists take responsibility for curation
 - Rapid data entry
 - Attractive UI
 - Hide complexity
- Gather structured information efficiently
 - Rich fact-gathering
- Centralise metadata gathering
 - De-duplication
 - Spreadsheets everywhere!

“Deckard”

- Off-the-shelf metadata management products investigated
 - Cost
 - Functionality
 - Technical integration
- In-house prototype project
 - People and Organisations
 - Searchable
 - Bulk upload
 - Candidate entity queue

Images removed for copyright reasons

A developer point of view

The Deckard project presented one main challenge.

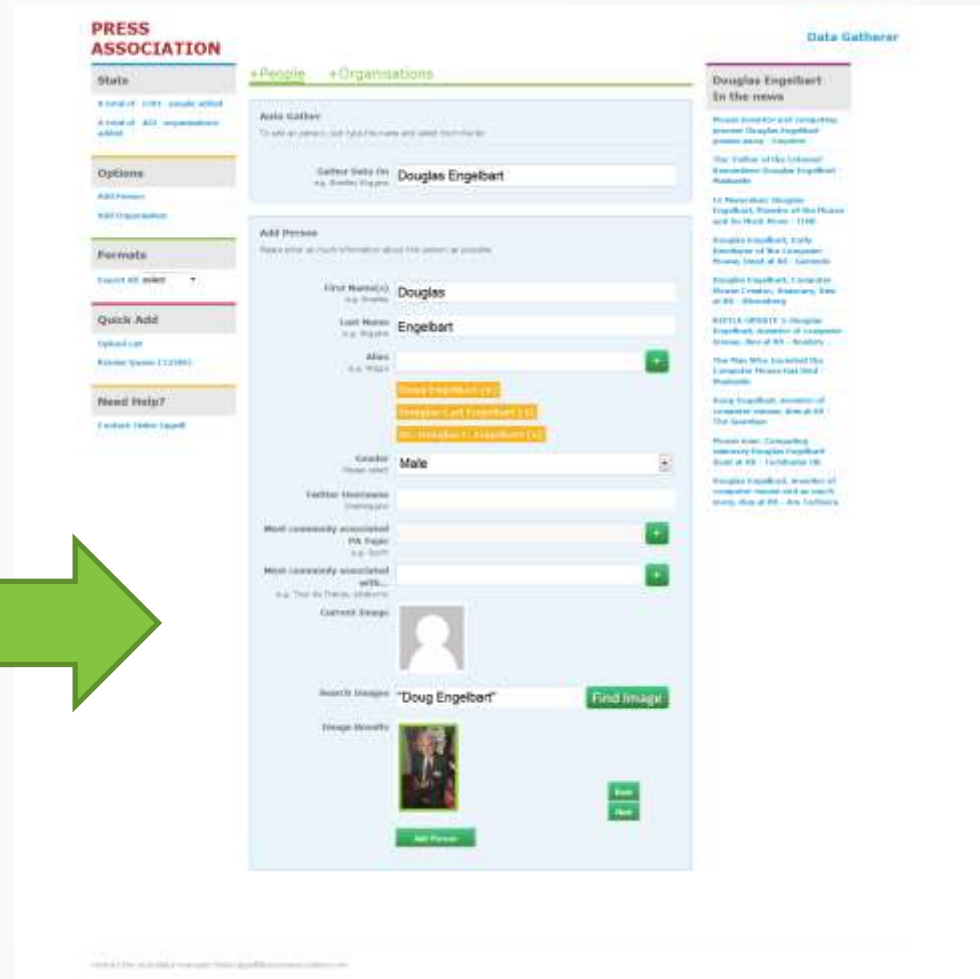
- To make the user want to use Deckard!

Which in turn unfortunately lead to a number of technical challenges!

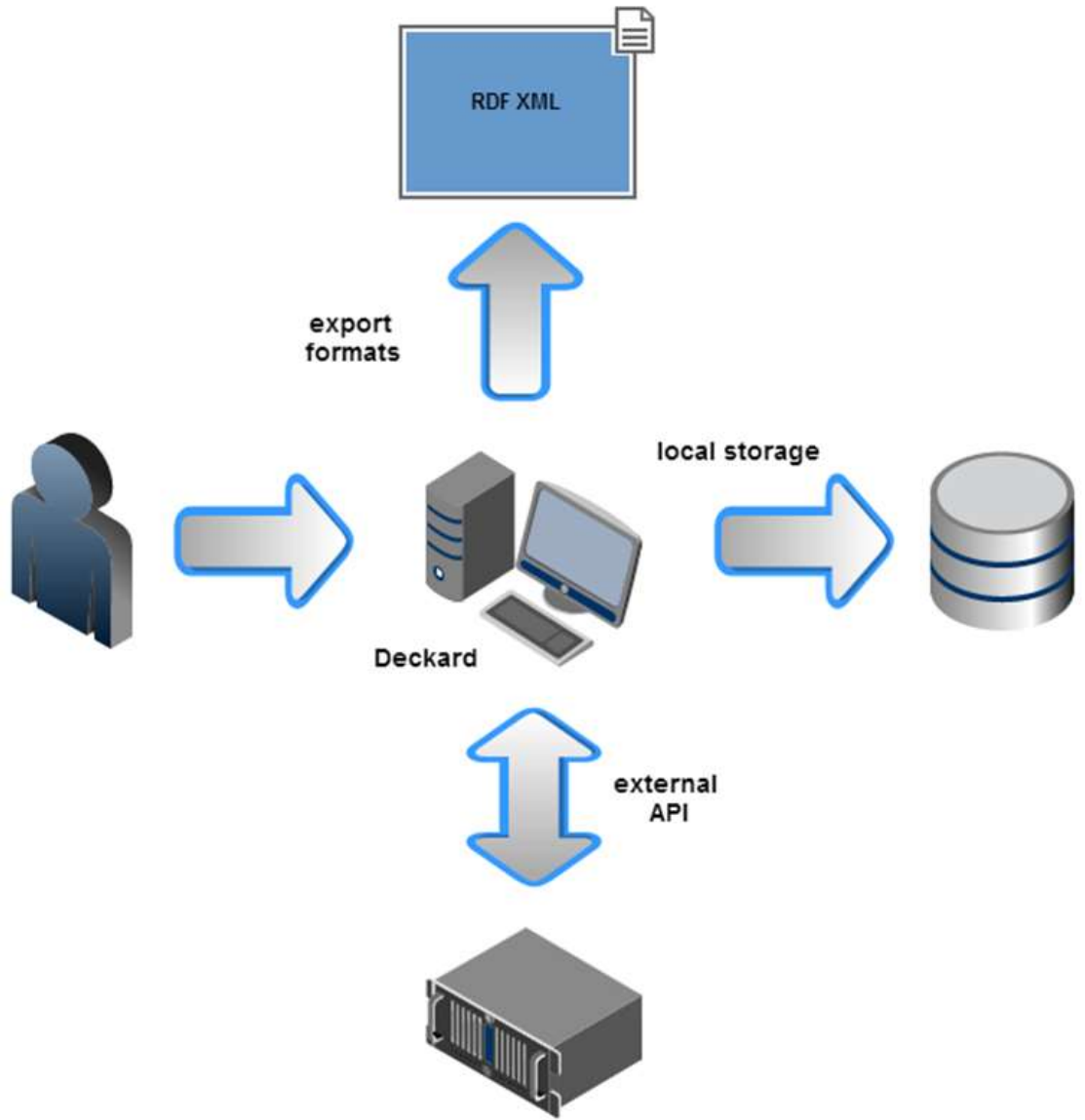
- Using external sources enhance manually inputted data
- Simple methods to edit an entity
- Immediate and rewarding feedback to users
- Outputs in various formats useful to the user and other parts of the business

Design

- As fun to use as possible
- A pleasant experience
- The least amount of clicks
- Maximum collection of data



Architecture



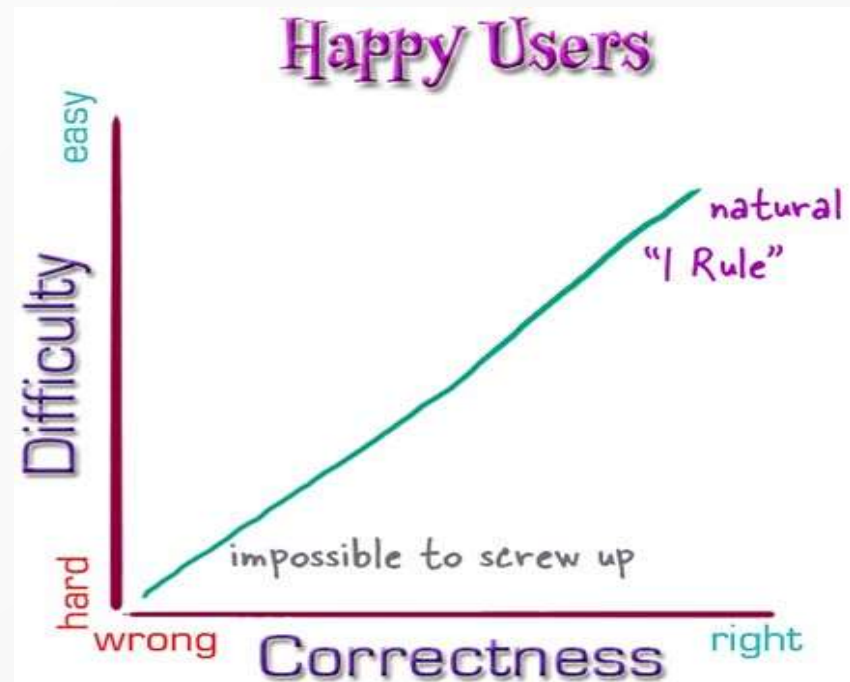
The Front End

The jQuery library used extensively to achieve the following:

- Dynamic interaction with data
- Increased speed with fewer page reloads
- Immediate feedback to the user

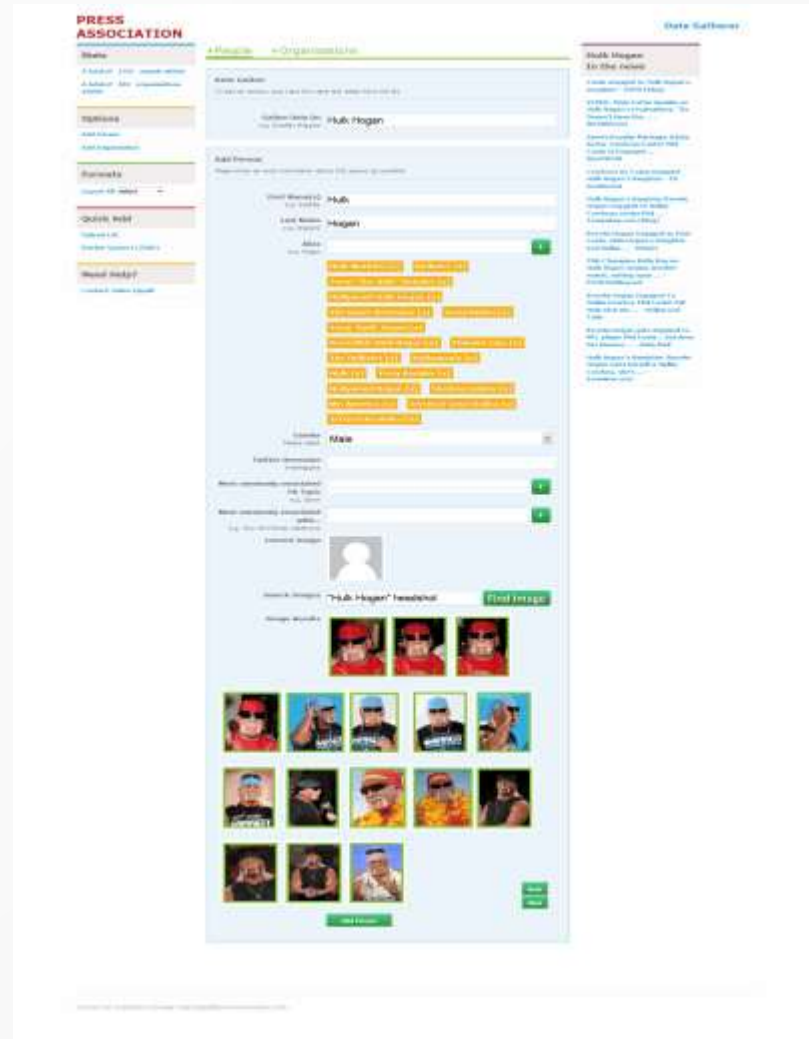
All the features users love!

Make the user happy,
keep them on Deckard,
feeding in the data!



The Front End - Examples

- Freebase search auto suggest
- Dynamic generation of PA image search results
- Data containers created and removed on the fly
- Updating side bars with relevant news and tweets



The Back End

A far more boring place to be!

In lines of code this is quite a small project so it was decided to develop it in PHP and MySQL

- Specifically designed for web applications
- Simple code language
- Respond quickly to changing specifications
- Easy to change data structures
- Quick deployment
- Testing frameworks – PHPUnit
- Many free to use PHP libraries available e.g. twitter, freebase

APIs

- Back end APIs

- Used to discover data
 - Freebase – entity discovery
 - Geonames – silent API
 - PA Images Search API

- Front end APIs

- Used to populate sidebars and enhance user experience
 - Wikipedia - Info box
 - Twitter API – latest tweets
 - Google News – recent news

What about the poor user?

So after all this back end crunching and processing what does the user get for all their efforts

PRESS ASSOCIATION

Stats

A total of 1301 people added

A total of 403 organisations added

Options

[Add Person](#)

[Add Organisation](#)

Formats

Export All

This Person

Quick Add

[Upload List](#)

[Review Queue \(12306\)](#)

Need Help?

Contact: Helen Lloyd

+People +Organisations



Patrick Moore

Sir Patrick Alfred Caldwell-Moore, CBE, FRS, FRAS (born 4 March 1923) is a British amateur astronomer who has attained prominent status in astronomy as a writer, researcher, radio commentator and television presenter, and who is credited as having done more than any other person to raise the profile of astronomy among the British general public. He is a former president of the British Astronomical Association, co-founder and former president of the Society for Popular Astronomy (SPA), author of over 70 books on astronomy, and presenter of the world's longest-running television series with the same original presenter, *The Sky at Night* on the BBC. As an amateur astronomer, he became known as a specialist on observing the Moon and creating the Caldwell catalogue. Idiosyncrasies such as his rapid diction and monocle have made him a popular and instantly recognisable figure on British television. Moore is also a self-taught xylophone and piano player and an accomplished composer, and a former amateur cricketer, golfer, and chess player. In addition to his many popular science books, he has written numerous works of fiction, and was the presenter of the 1990s TV series *GamesMaster*. An

Gender: male
Born: 4th of March 1923
Place Of Birth: Kinner
Died: 2012-12-09
Common Aliases: Sir Patrick Alfred Caldwell-Moore
Sir Patrick Moore
Alfred Patrick Caldwell-Moore



[View Larger Map](#)

Professions

Astronomer
Presenter
Writer

Associations

Astronomy
British Broadcasting Corporation
The Sky at Night

PA Topics

MEDIA
SCIENCE

More Links

<http://en.wikipedia.org/wiki/index.htm?curid=189654>
<http://openlibrary.org/authors/OL28366A>
<http://www.daylife.com/topic/66a959afC26d>
<http://www.tvrage.com/person/id-21793>
<http://id.loc.gov/authorities/names/n79082293.html>
<http://viaf.org/viaf/32062500/>
<http://de.wikipedia.org/wiki/index.htm?curid=271448>
<http://es.wikipedia.org/wiki/index.htm?curid=699601>
<http://it.wikipedia.org/wiki/index.htm?curid=1837319>
<http://fr.wikipedia.org/wiki/index.htm?curid=4507286>

contact the metadata manager helen.lloyd@pressassociation.com

Data Gatherer

Patrick Moore In the news

[Stars pay tribute to Sir Patrick Moore, the man who took us into space - The Independent](#)

[Sir Patrick Moore, 89, Dies At Home - Sky News](#)

[Sir Patrick Moore: the eccentric amateur who became a TV star | Science | The ... - The Guardian](#)

[Sir Patrick Moore dies aged 89 - video | Science | guardian.co.uk - The Guardian](#)

[Sir Patrick Moore: Astronomer who presented 'The Sky At Night' for over half a ... - The Independent](#)

[Patrick Moore tributes pour in after astronomer dies aged 89 | UK news - The Guardian](#)

[Sir Patrick Moore: Chris Laskett's tribute - BBC News](#)

[Sir Patrick Moore: Your memories - BBC News](#)

[Sir Patrick Moore - Telegraph.co.uk](#)

[Sir Patrick Moore dies aged 89 - Telegraph.co.uk](#)

Export Formats

Data can be exported in the following formats:

- RDF XML
- Turtle
- Triples
- JSON

```
01.         @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
02. @prefix pns: <http://data.press.net/ontology/stuff/> .
03. @prefix pni: <http://data.press.net/ontology/identifier/> .
04. @prefix foaf: <http://xmlns.com/foaf/0.1/> .
05. @prefix time: <http://www.w3.org/2006/time#> .
06. @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
07.
08. <http://data.press.net/stuff/b75a6a0f-3f9a-4639-8739-ea54ff06b43f>
09.   a pns:Person ;
10.   pni:hasIdentifier <http://rdf.freebase.com/ns/en.idris_elba> ;
11.   foaf:givenName "Idris"@en ;
12.   foaf:familyName "Elba"@en ;
13.   pns:name "Idris Elba"@en ;
14.   pns:alias "Idrissa Akuna Elba"@en, "DJ Big Driis"@en, "Big Driis The Londoner"@en, "E.
15.   foaf:gender "male" ;
16.   pns:placeOfBirth <http://sws.geonames.org/2646476/> ;
17.   pns:dateOfBirth [ time:inXSDDateTime "1972-09-06T00:00:00Z"^^xsd:dateTime ] ;
18.   pns:notablyAssociatedWith <http://rdf.freebase.com/ns/en/the_wire>, <http://rdf.freeb
19.   pns:preferredImage <http://data.press.net/assets/i/2.11643964> .
20.
21. <http://rdf.freebase.com/ns/en.idris_elba>
22.   a pni:Identifier ;
23.   pni:authority <urn:freebase.com> .
24.
25.
26.
27. <http://data.press.net/stuff/4d76a14d-e7e6-47a3-9f2a-a587ead137f1>
28.   a pns:Person ;
29.   pni:hasIdentifier <http://rdf.freebase.com/ns/en.ukasz_kubot> ;
30.   foaf:givenName "Lukasz"@en ;
31.   foaf:familyName "Kubot"@en ;
32.   pns:name "Lukasz Kubot"@en ;
33.   pns:alias "\u00C5\u0081lukasz Kubot"@en ;
34.   foaf:gender "male" ;
35.   pns:placeOfBirth <http://sws.geonames.org/3102987/> ;
36.   pns:dateOfBirth [ time:inXSDDateTime "1982-05-16T00:00:00Z"^^xsd:dateTime ] ;
37.   pns:notablyAssociatedWith <http://rdf.freebase.com/ns/m/0k8fkv2>, <http://rdf.freebas
38.   pns:preferredImage <http://data.press.net/assets/i/2.16957560> .
39.
40. <http://rdf.freebase.com/ns/en.ukasz_kubot>
41.   a pni:Identifier ;
42.   pni:authority <urn:freebase.com> .
43.
44.
45.
46. <http://data.press.net/stuff/377345e5-20bd-4058-8e28-e2bbece2de62>
47.   a pns:Person ;
48.   pni:hasIdentifier <http://rdf.freebase.com/ns/en.nicole_farhi> ;
49.   foaf:givenName "Nicole"@en ;
50.   foaf:familyName "Farhi"@en ;
51.   pns:name "Nicole Farhi"@en ;
52.   pns:alias "Lady Hare"@en ;
53.   foaf:gender "female" ;
54.   pns:placeOfBirth <http://sws.geonames.org/2990440/> ;
55.   pns:dateOfBirth [ time:inXSDDateTime "1946-07-25T00:00:00Z"^^xsd:dateTime ] ;
56.   pns:notablyAssociatedWith <http://rdf.freebase.com/ns/user/jihnken/fashion/fashion_re
57.   pns:preferredImage <http://data.press.net/assets/i/2.5326235> .
58.
```


Future Development

- Move away from MySQL and write data straight to a Mark Logic database before that data will be exported to the triple store
- Addition of more entities e.g. events
- Pull in more data from more external sources
 - IMDB
 - Facebook
- Pull in more data from internal sources
 - PA News Feeds
 - PA Video
 - PA Galleries

- Back to Helen 😊

Business engagement

- Small, targeted data curation pilots
- Getting feedback on UI
- Use familiar terminology e.g. PA Topics
- Workshops and smaller progress meetings

What worked well

- Demonstrable and iterative progress
- Clean UI focussed around our requirements
- Easy way of building up “newsworthy” data set
- Collaborative development process
- Stimulated interest in semantics

Challenges

- Wider business changes
- Resource constraints
- Tagging seen as 'more work'
- Building trust in the data
- Chicken and egg of metadata benefits
- Tool not production-ready...yet
- Workflow v need to break news

Data considerations # 1

- Disambiguation
 - Validation not fuzzy yet e.g. salutations, middle initials
 - Transliterations
 - Punctuation
 - Jonathan Davies, James Wharton, FSA, Liberty
- Adding related concepts
 - e.g. Roles, organisations, spouses, scandals, films
- Ongoing maintenance and updates
 - e.g. Deaths, marriages, honours, job moves, name changes
 - Mrs Thatcher, Jose Mourinho, EE, Lily Allen, Heathrow, Sir Tony Robinson, News UK
- Provenance and quality of third party data

Data considerations #2

- Topicality and building up the semantic graph
 - Leveson, horsemeat, recession, Operation Yewtree, MPs
- Planning around scheduled events
 - Awards, sports tournaments, TV shows

Images removed for copyright reason





Thank you

helen.lippell@pressassociation.com

elizabeth.perreau@pressassociation.com