University of BRISTOL

SubSift: a novel application of the vector space model to support the academic research process

Simon Price
Institute for Learning and Research Technology

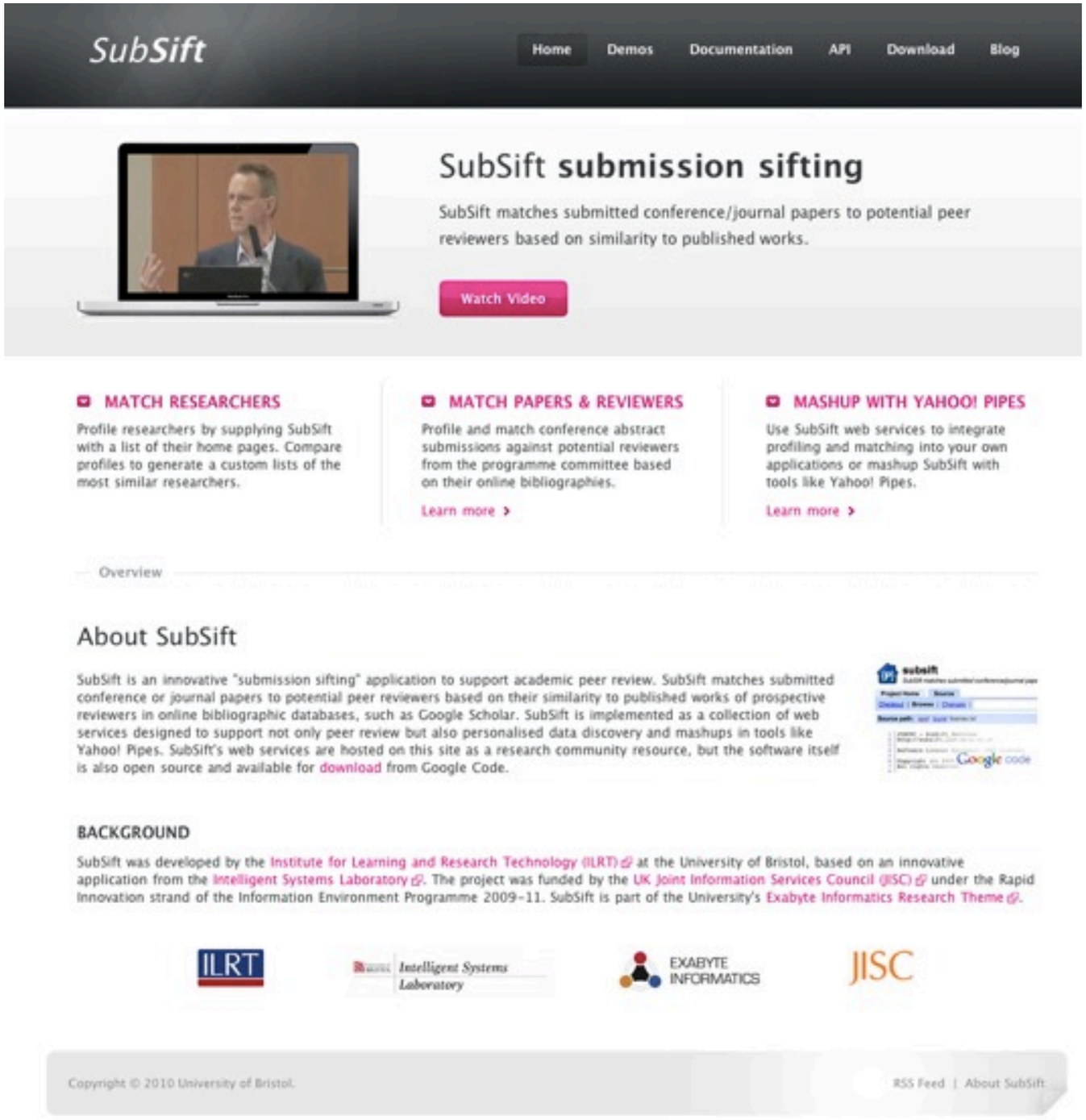
Peter A. Flach and Sebastian Spiegler
Intelligent Systems Laboratory

SubSift

- SubSift is a prototype application to support academic peer review.

- SubSift matches submitted conference/journal papers to potential peer reviewers based on similarity to published works.

- Website:
  http://subsift.ilrt.bris.ac.uk



Contribution of this work

- Innovative application of established theory
- Open Source software
- Hosted web services
- Example applications

Outline of this paper

1. Motivation and Implementation
2. Background Theory
   - Vector Space Model
   - Representational State Transfer (REST)
3. SubSift Web Services
4. Applications

1. Motivation and Implementation

Motivation: KDD'09 review process

Peter Flach was programme committee (PC) co-chair

- 500+ papers and 200+ PC members
- Idea: streamline the paper bidding and allocation process
- Software developed to do this - *part of which we named SubSift*
  - bid initialisation (3=want to review, ..., 0=do not want)
  - papers ranked for each PC member
  - PC members ranked for each paper

Further details:

Peter A. Flach, Sebastian Spiegler, Bruno Golénia, Simon Price,
John Guiver Ralf, Herbrich, Thore Graepel, and Mohammed J. Zaki.
Novel tools to streamline the conference review process: Experiences from SIGKDD'09
SIGKDD Explorations, 11(2):63–67, December 2009

Evaluating SubSift at KDD'09

- Precision and recall:
  - 88% median precision
    (non-zero actual bids among non-zero predicted bids)
  - 80% median recall
    (non-zero predicted bids among non-zero actual bids)

- User feedback:

  "…as I go thru my paper assignments, I am extremely impressed by quality of your initial automated assignment!"

  Gregory Piatetsky (KDD'09 reviewer)

Implementation

- Project to repackage SubSift as web services

ILRT → Submitted bid: ***SubSift Services*** → JISC

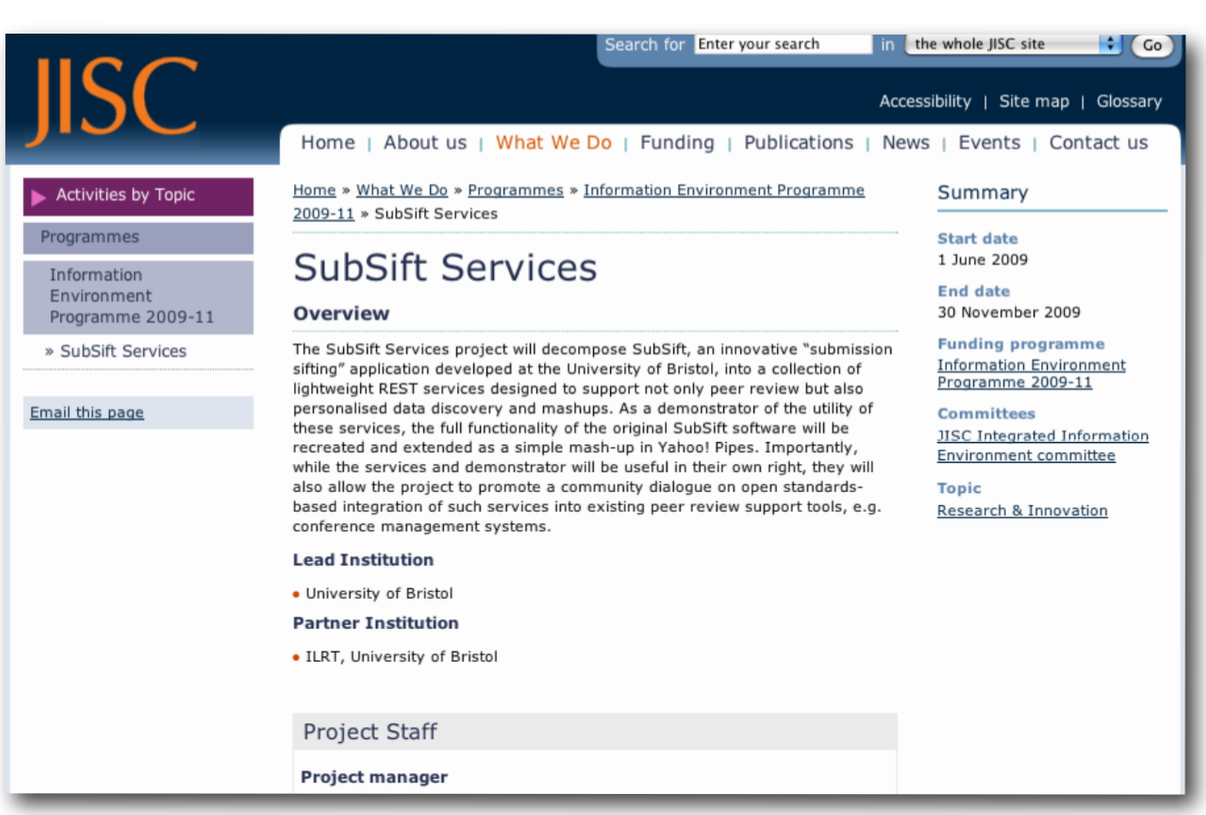
"Rapid Innovation" call under the JISC Information Environment Programme

Implementation

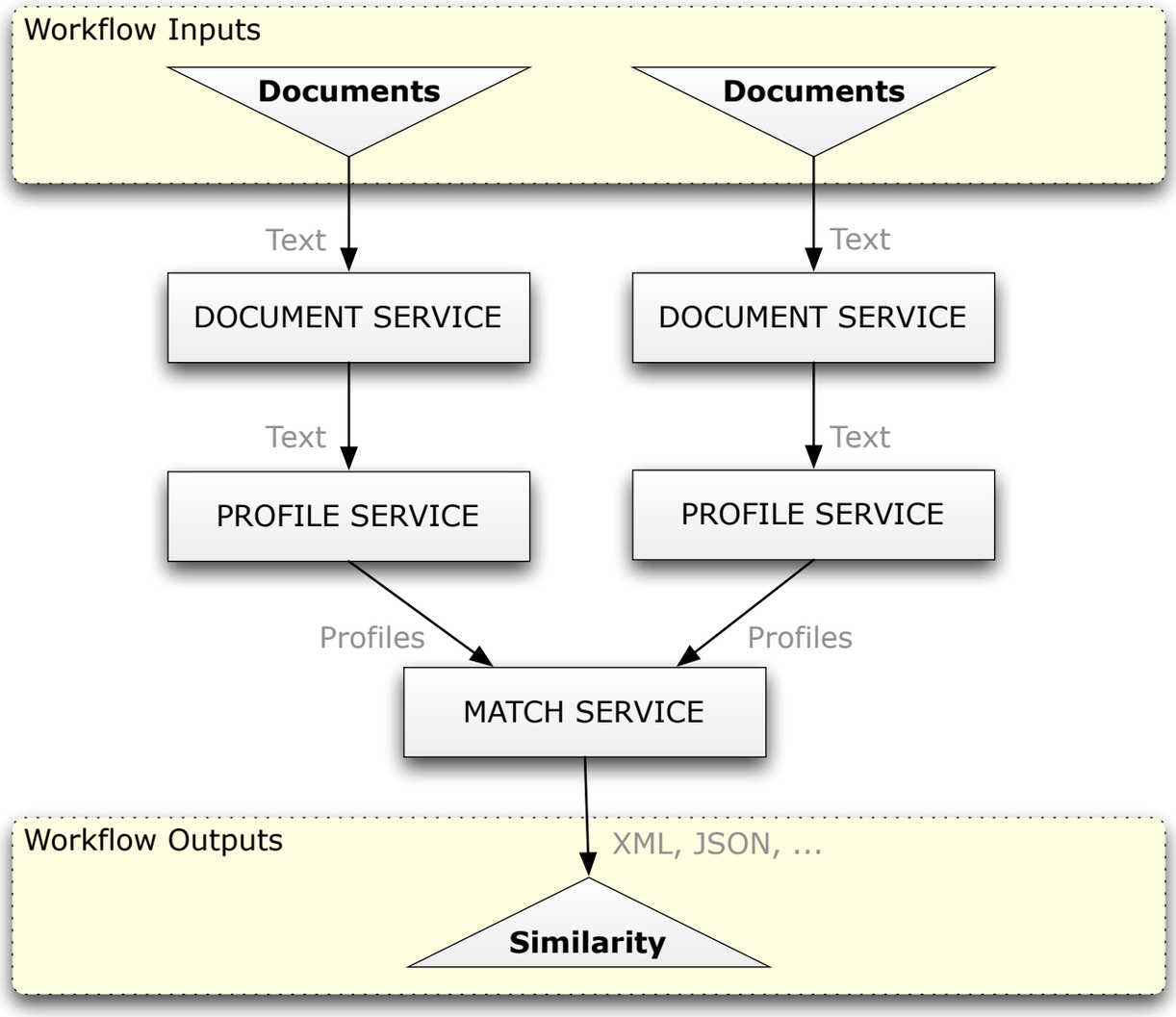
- Project to repackage SubSift as web services

ILRT → Submitted bid: ***SubSift Services*** → JISC

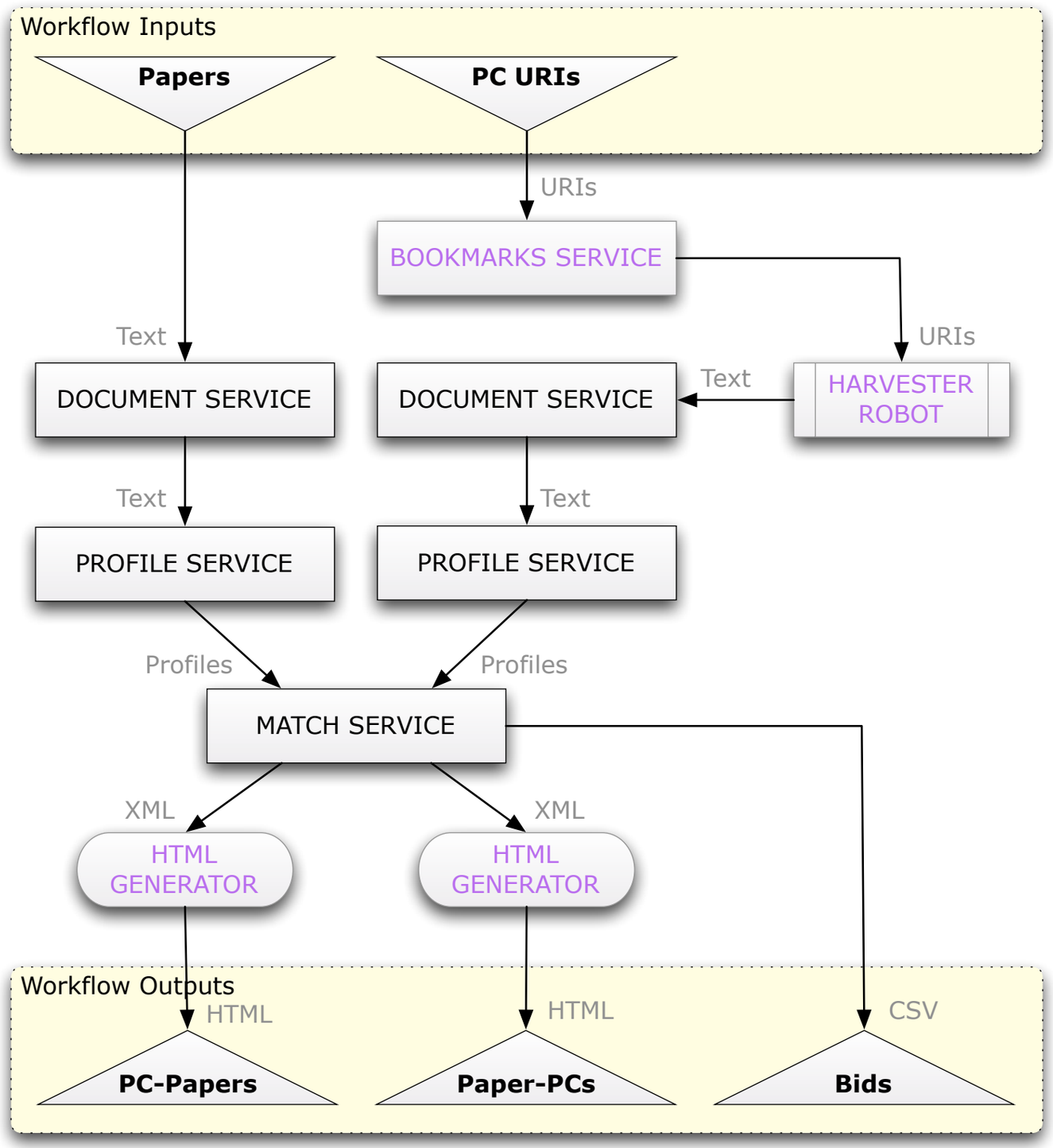
"Rapid Innovation" call under the JISC Information Environment Programme



SubSift Services



SubSift Services



2. Background Theory

University of BRISTOL

Vector Space Model (from Information Retrieval)

For a query (q), rank the documents (d_j) in collection (D) by descending similarity to the query.

Vector Space Model consists of:

- *bag-of-words* representation
- cosine similarity
- tf-idf weighting

Vector Space Model: *bag-of-words* representation

no. terms in each abstract

	intelligence	learning	logic	machine
abstract 1	0	2	0	1
abstract 2	3	0	5	0
abstract 3	2	1	0	2

no. terms in DBLP author page of each PC member

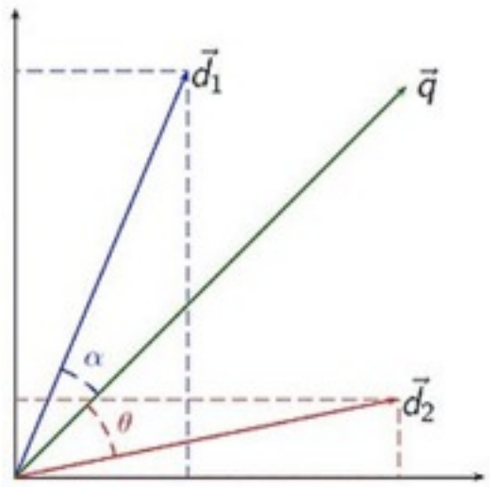
	intelligence	learning	logic	machine
pc member 1	10	70	20	0
pc member 2	0	70	5	99
pc member 3	30	70	0	0

Vector Space Model: cosine similarity

Query and document similar if angle θ between their vectors is small.

$$\textbf{similarity}_{\text{cosine}}(\vec{q}, \vec{d}) = \cos(\theta) = \frac{\vec{q} \cdot \vec{d}}{||\vec{q}|| \cdot ||\vec{d}||}$$

- $d \in D$ – document represented as multiset of terms (*bag-of-words*).
- $\vec{d}$ – document vector in the *vector space* defined by vocabulary of D.
- $\vec{q}$ – query vector in the same vector space as $\vec{d}$.



Vector Space Model: tf-idf weighting

Normalise term counts within document and penalise common terms in D.

$$\mathrm{tf}_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \qquad \mathrm{idf}_i = \log_2 \left(\frac{|D|}{\mathrm{df}_i} \right), \qquad \text{tf-idf}_{ij} = \mathrm{tf}_{ij} \times \mathrm{idf}_j$$

- tf_{ij} is *term frequency* of term t_i in the document d_j.
- n_{ij} is *term count*, the number of times term t_i occurs in the document d_j.
- df_i is *document frequency* of term t_i is the number of documents in D in which term t_i occurs.

Vector Space Model: tf-idf weighting

Normalise term counts within document and penalise common terms in D.

$$\mathrm{tf}_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \qquad \mathrm{idf}_i = \log_2 \left(\frac{|D|}{\mathrm{df}_i} \right), \qquad \text{tf-idf}_{ij} = \mathrm{tf}_{ij} \times \mathrm{idf}_j$$

- tf_{ij} is *term frequency* of term t_i in the document d_j.
- n_{ij} is *term count*, the number of times term t_i occurs in the document d_j.
- df_i is *document frequency* of term t_i is the number of documents in D in which term t_i occurs.

	intelligence	learning	logic	machine
pc member 1	10	70	20	0
pc member 2	0	70	5	99
pc member 3	30	70	0	0

Representational State Transfer (REST)

REST is a design pattern for **web services** based on HTTP using its familiar URIs, requests, responses, authentication, etc.

"RESTful" web services:

- URIs to represent resources
- HTTP POST/GET/PUT/DELETE correspond to usual Create/Read/Update/Delete (CRUD) operations
- Response formats typically include: XML, JSON, CSV

SubSift Web Services

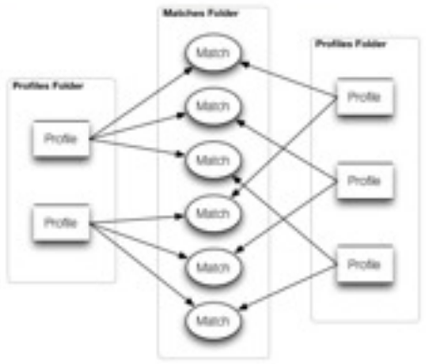
1. Motivation and Implementation
2. Background Theory
3. SubSift Web Services
4. Applications

University of BRISTOL

SubSift REST API

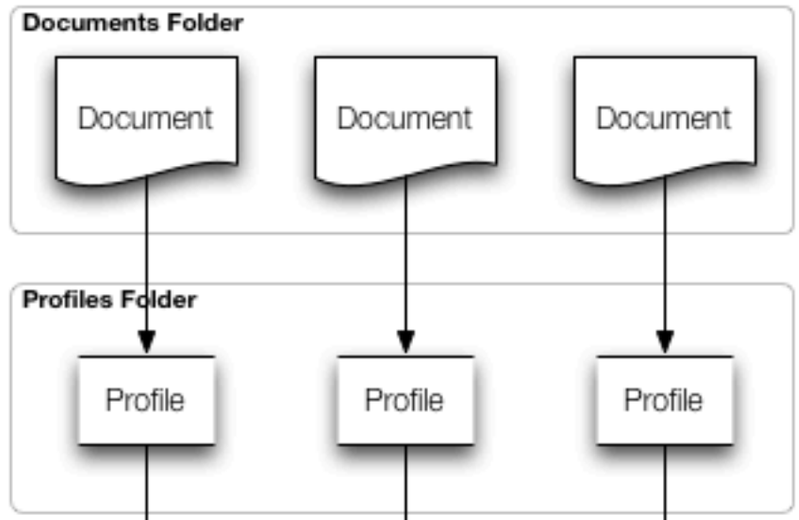
3. Matches

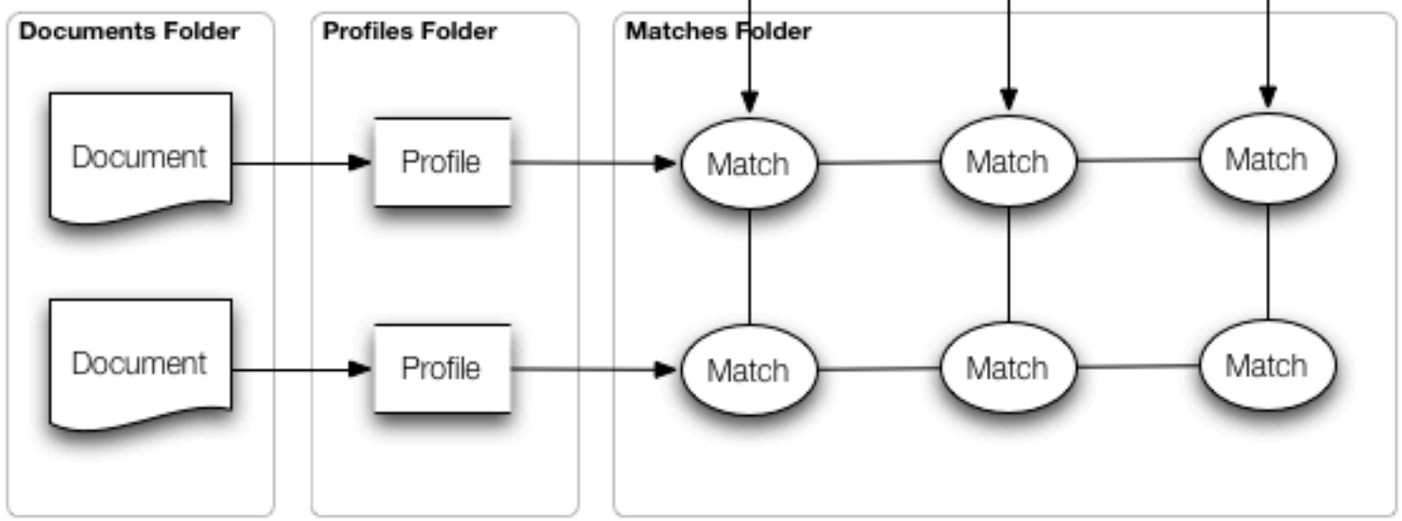
In SubSift, a match item is a similarity score (and supporting statistics) representing how alike a specific pair of profile items are. Each matches folder is a container to hold a list of match items. A matches folder is created by analysing every pairing of profile items drawn from a pair of profiles folders. Each match item scores the similarity of a single profile from the first profiles folder against every profile from the second profiles folder. A typical usage of such a comparison is to match submitted conference abstracts with the bibliography pages of programme committee members in order to rank potential reviewers for each paper and visa versa.



3.1 MATCHES FOLDERS

API Method	HTTP	URI Schema	Parameters
matches list	GET	/:user_id/matches	sort, full
matches show	GET	/:user_id/matches/:folder_id	sort, full
matches exists	HEAD	/:user_id/matches/:folder_id	
matches create	POST	/:user_id/matches/:folder_id/profiles/:profiles_id1/with/:profiles_id2	description, mode, limit, threshold, sort, full
	POST	/:user_id/matches/:folder_id	profiles_id1, profiles_id2, description, mode, limit, threshold, sort, full
matches update	PUT	/:user_id/matches/:folder_id/profiles/:profiles_id1/with/:profiles_id2	description, mode, limit, threshold, sort, full





Applications

1. Motivation and Implementation
2. Background Theory
3. SubSift Web Services
4. Applications

University of BRISTOL

SubSift has been used for...

Profiling a research group by its publications

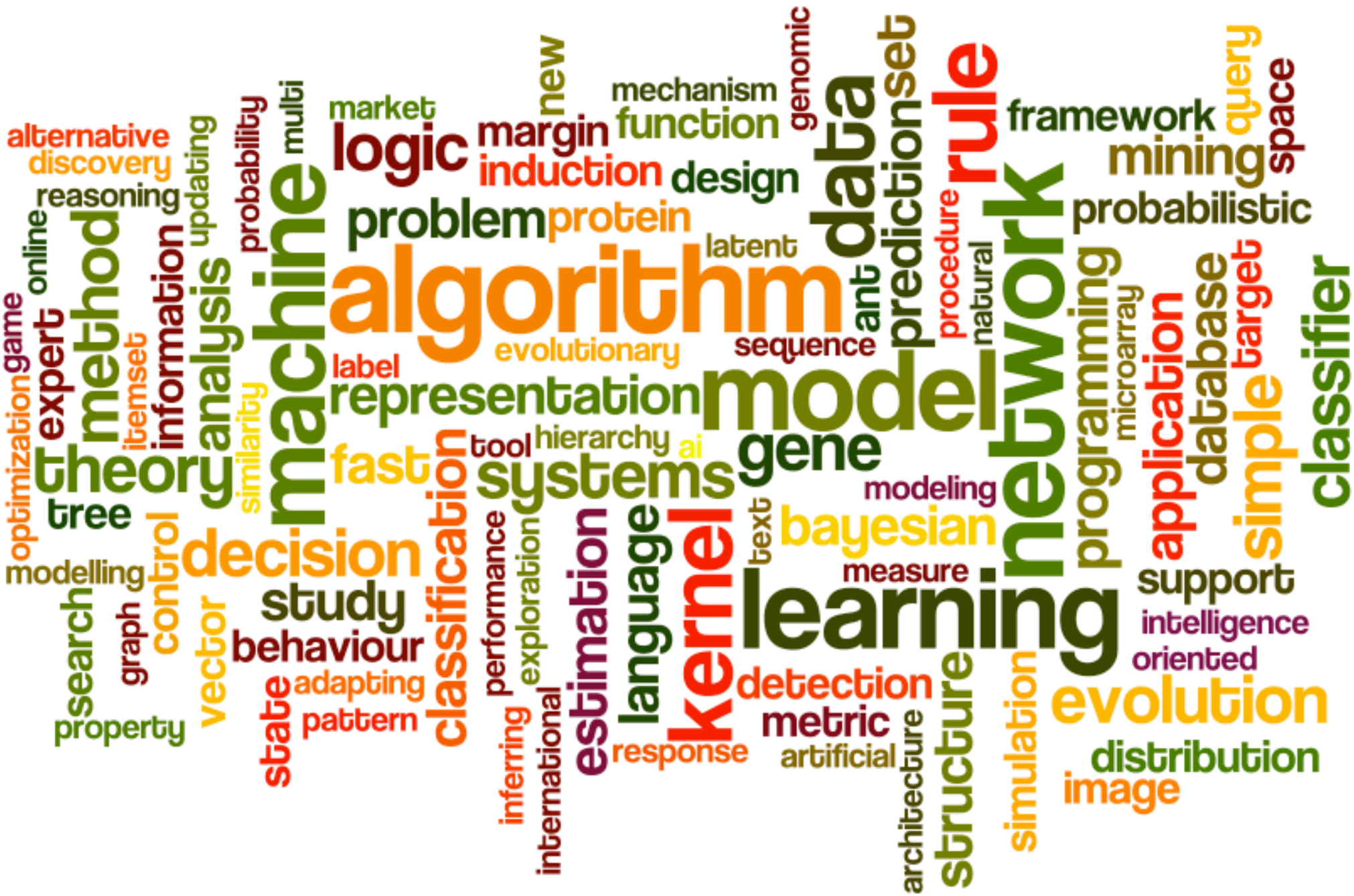


Diagram produced in Wordle using SubSift profile data

Finding an expert

ILRT Matcher

Enter a title, abstract or text of a paper and click Submit to compare against a pre-defined set of profiles.

Text:

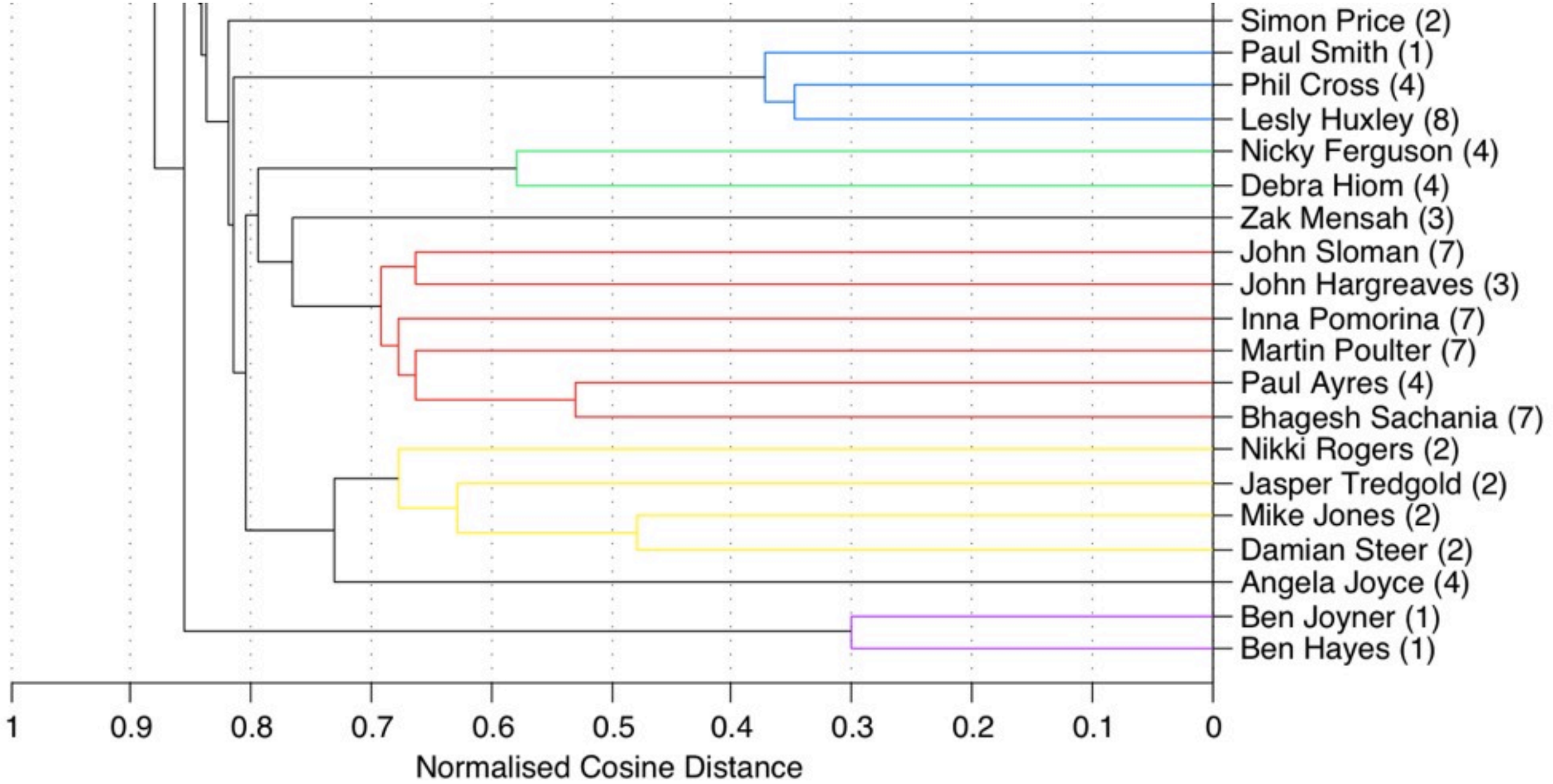
Semantic Web technologies have moved beyond the point of being promising futuristic technologies and demonstration projects, to being technologies in action in realistic contexts and conditions. Semantic Web applications are being developed for many aspects of scientific research, from experimental data management, discovery and retrieval, to analytic workflows, hypothesis development and testing, to research publishing and dissemination. This workshop intends to explore the questions that arise as Semantic Web applications are increasingly grounded within the actual lifecycle of scientific research, from observation and hypothesis formulation to publication, dissemination and criticism. We aim to bring together researchers across the disciplines, to discuss the use, development and embedding of these technologies in varied research domains and contexts. We will discuss the actuality of Semantic Web technologies in use and the emergent practices through which they are being developed and deployed. We aim to encourage vigorous discussion around aims, methods, applications and pragmatics. This workshop will look at the theoretical, methodological and pragmatic issues of grounding the development, deployment and evolution of ontologies and applications in Semantic e-

Submit

ILRT staff ranked by similarity to text

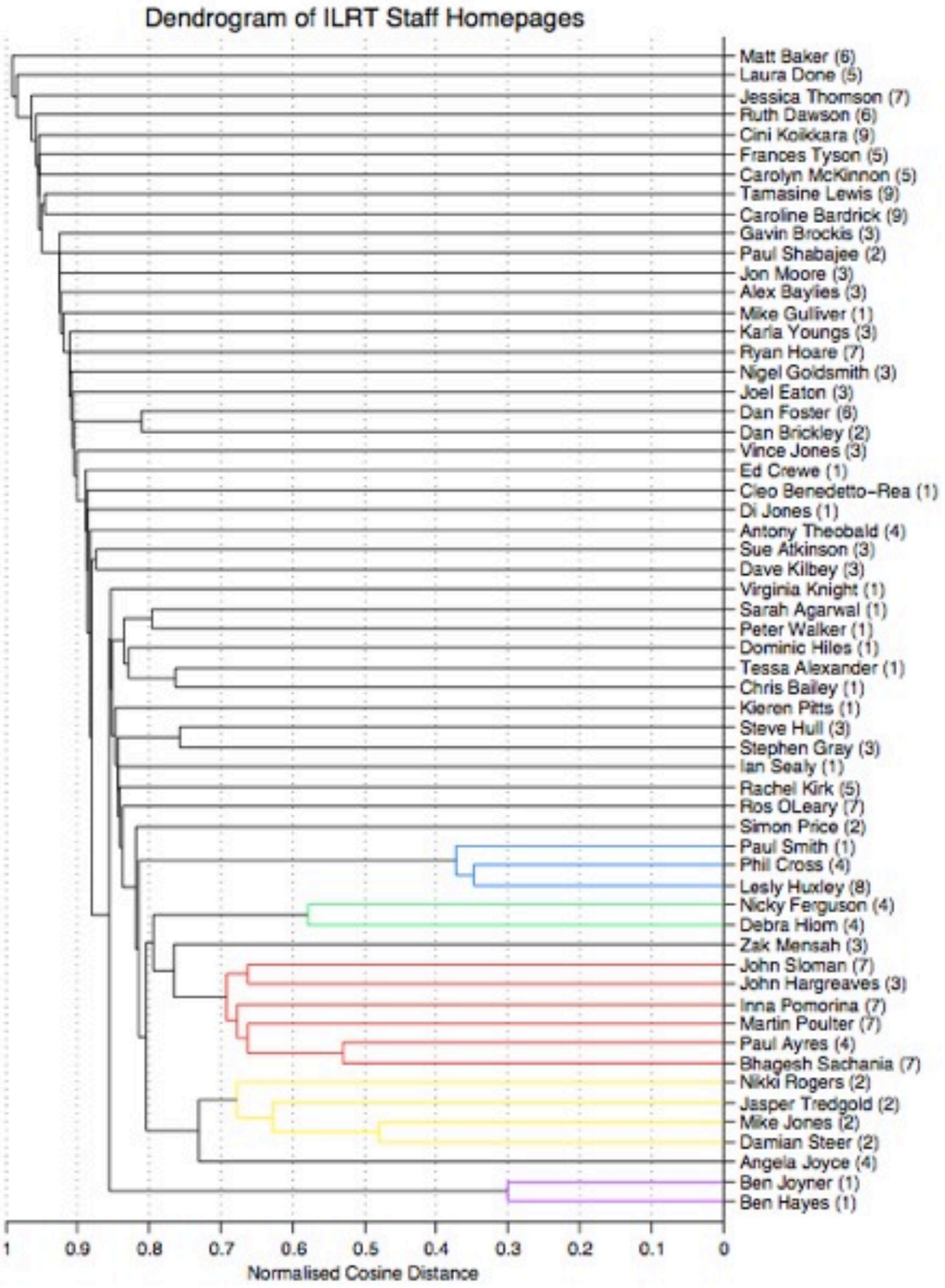
Damian Steer	0.142
Mike Jones	0.124
Nikki Rogers	0.095
Jasper Tredgold	0.072
Sarah Agarwal	0.059
Simon Price	0.057
Ben Joyner	0.045
Paul Shabajee	0.045
Chris Bailey	0.043

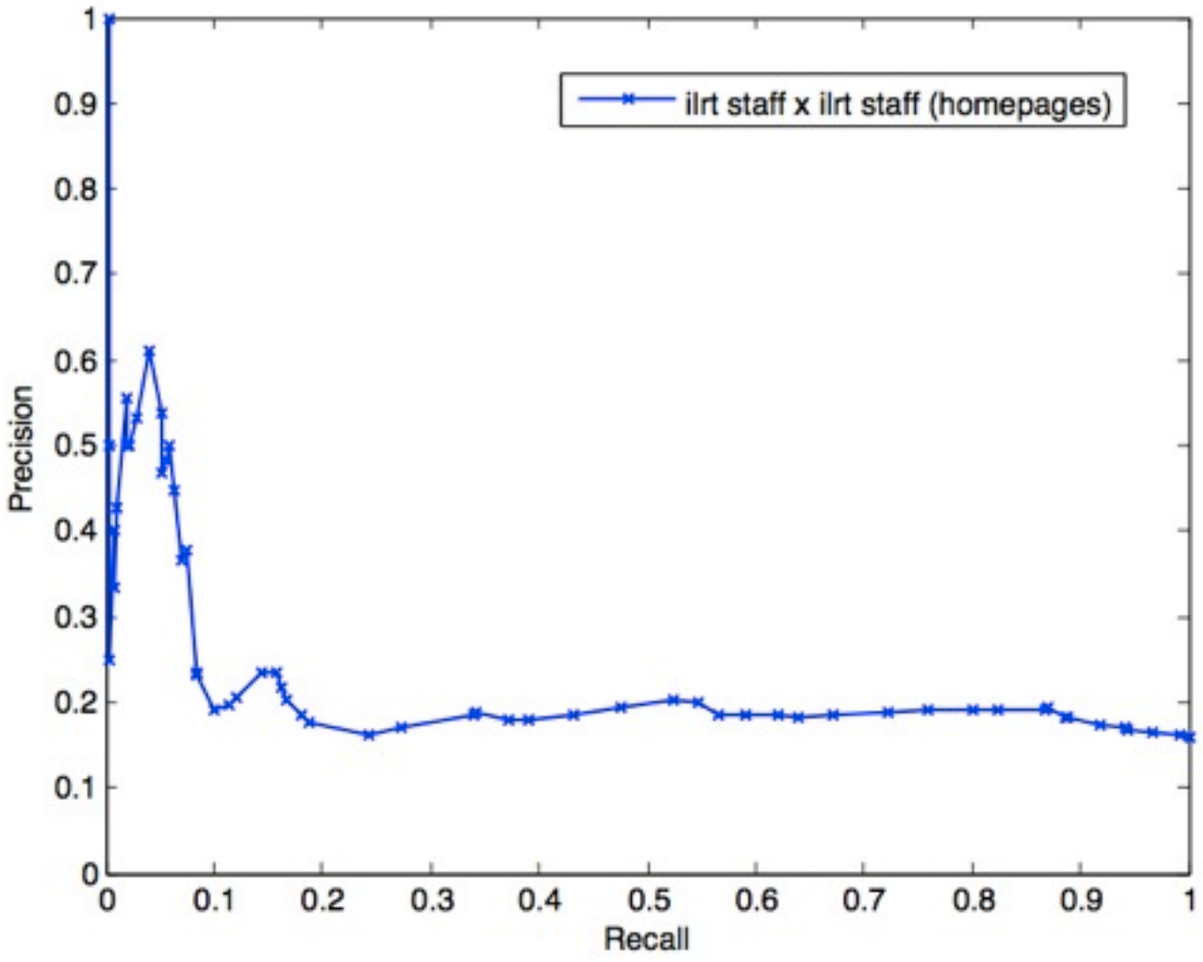
Clustering staff based on homepage similarity



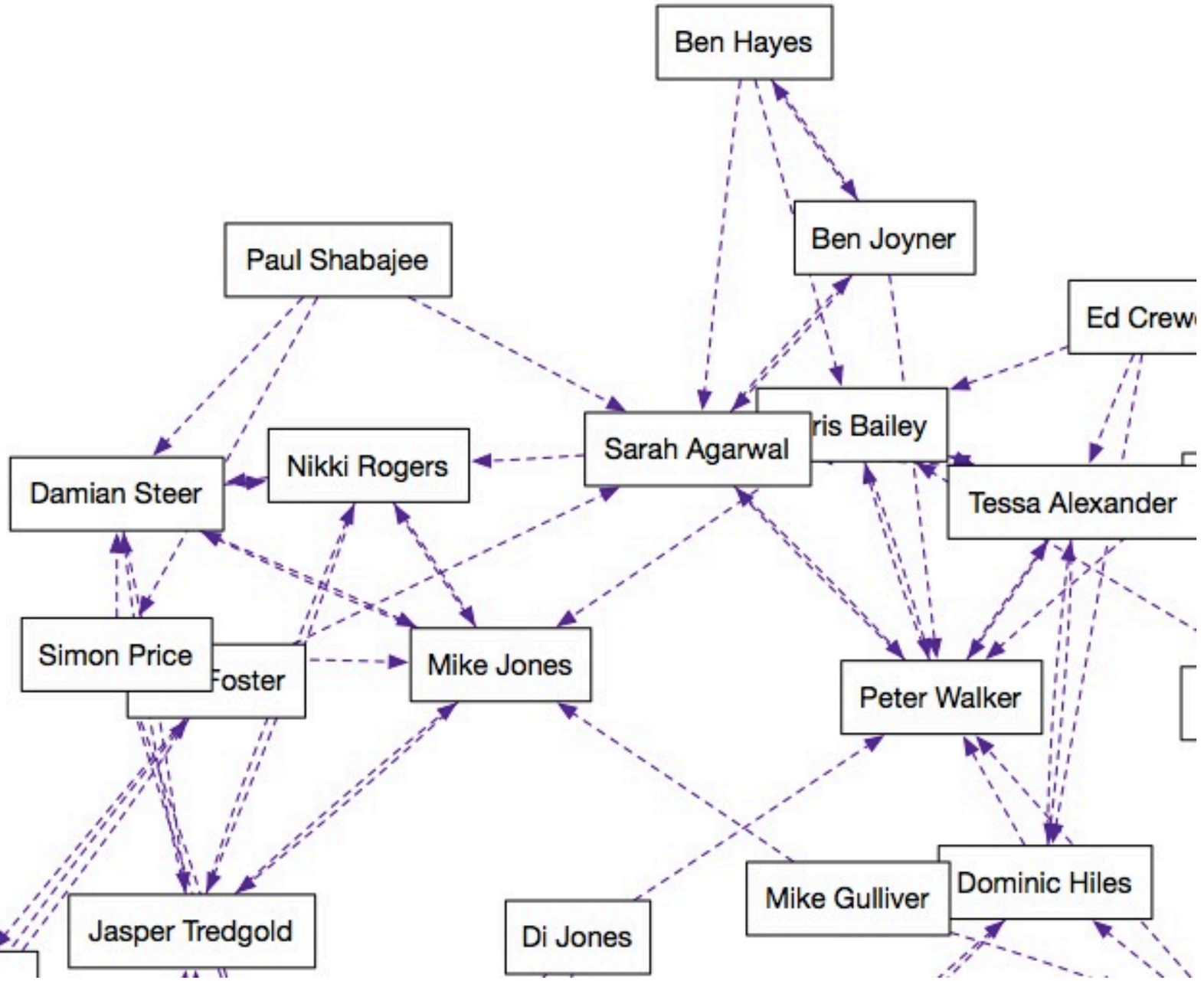
Dendrogram produced in Matlab from SubSift generated similarity matrix

Precision-recall at different thresholds





Similarity networks



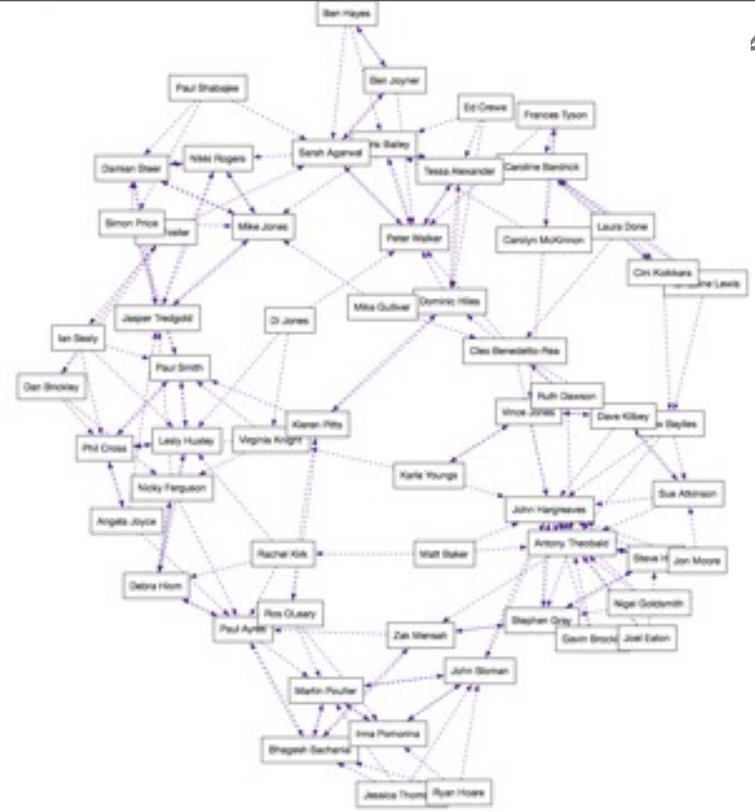
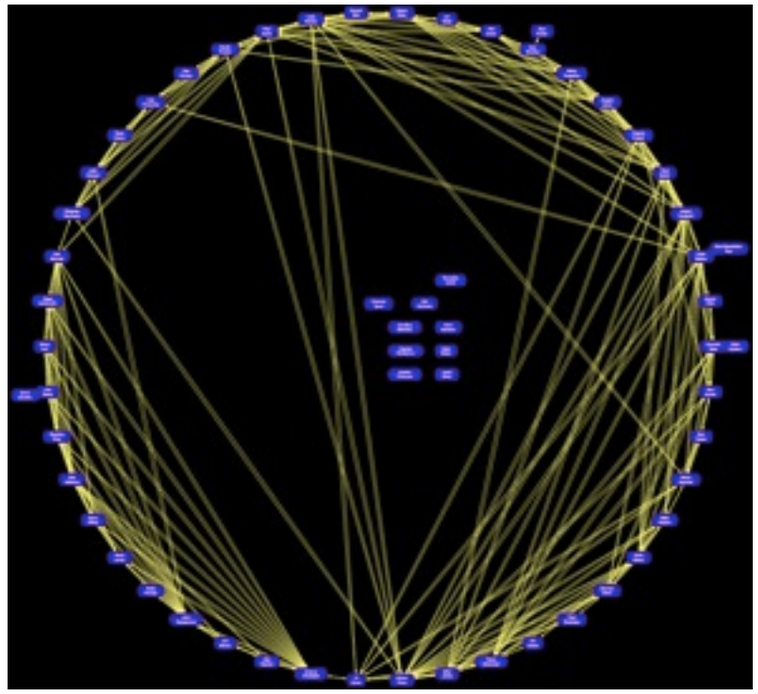


Diagram created by Graphvis from SubSift generated dot file

Connectivity



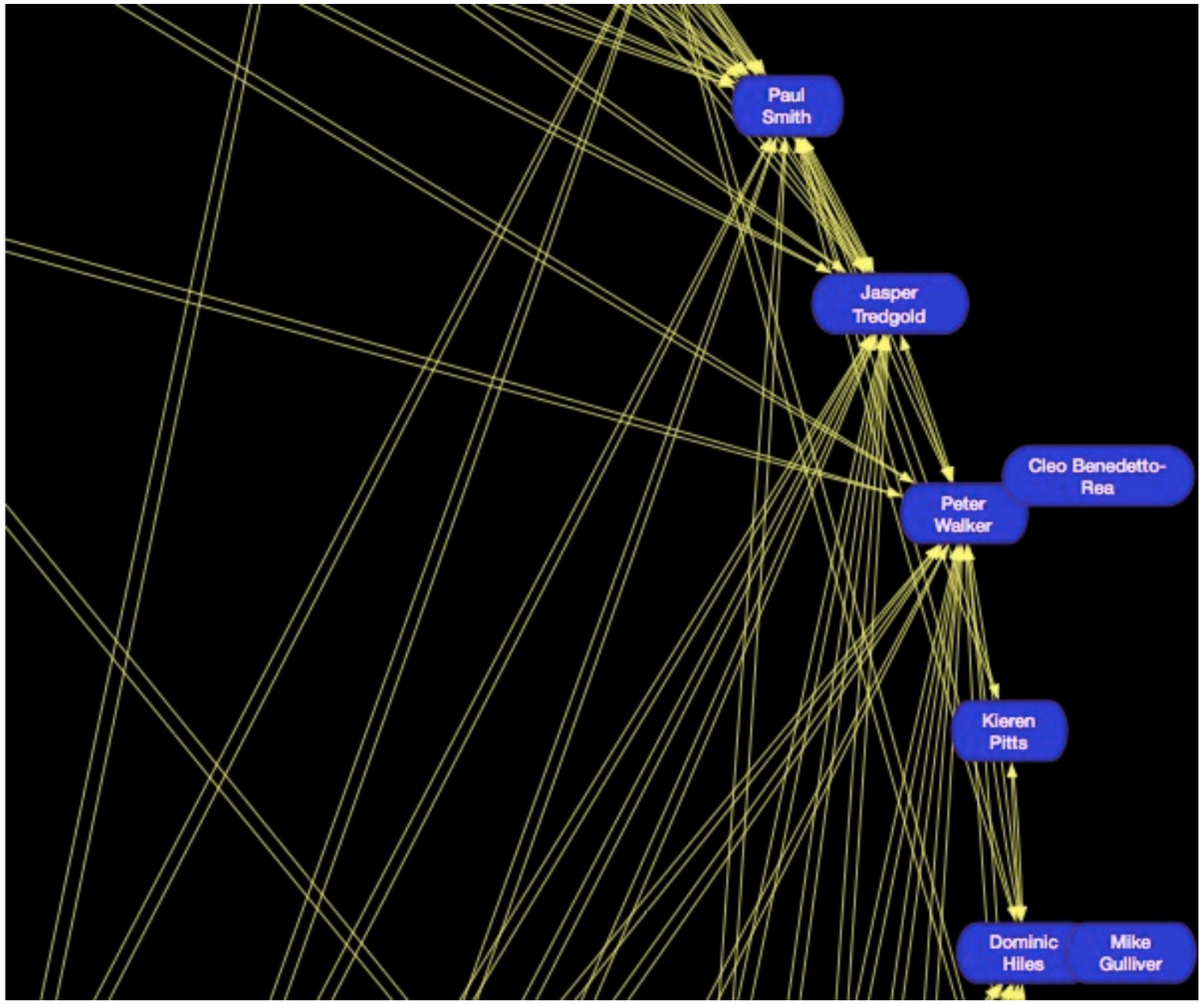


Diagram created by Graphvis from SubSift generated dot file

And finally...

University of BRISTOL

Conclusion

Repackaging SubSift as SubSift Services

- Created a more general purpose resource
- Potential applications outside of peer review domain

Publishing functionality as web services

- Similar approach may work for other research-produced applications