COMP38120 Workshop 1: What we mean by “Documents, Services and Data on the Web”

Jock McNaught, Riza Batista-Navarro, Norman Paton, Sandra Sampaio

The Staff

- In the order you encounter them today:
  - – Norman Paton (Laboratory, Services on the web).
  - – Jock McNaught (Documents on the web).
  - – Sandra Sampaio (Services on the web).
  - – Riza Batista-Navarro (Linked open data).

Workshop Outline

- Today's workshop seeks to give you a flavour of what to expect, with the following sections:
  - Overview of unit.
  - Documents on the web.
  - Services on the web.
  - Data on the web.
- This is a workshop, so there will be some up-front material (like this), and some hands-on activities.

Position in Workshop

- Overview of unit.
- Documents on the web.
- Services on the web.
- Data on the web.

Aim

- From the syllabus:
  - The aim of this course unit is to provide insights into and experience of techniques relating to documents, services and data on the web. The approach is that fundamental drivers, concepts and techniques for web documents, services and data are presented and discussed in workshop settings, and that a laboratory applies and evaluates the techniques in practice.
- The syllabus is at:
  - http://studentnet.cs.manchester.ac.uk/syllabus/index.php?code=COMP38120&year=2017

Things we don’t do

- We don’t do:
  - Web design.
  - Web application development.
  - Web data representation (XML, JSON, …).
  - Web servers.
  - Web standards.
  - ….
- But hopefully we have chosen a sensible and coherent collection of plausible topics, on which more later.

Structure

- Not absolutely standard stuff:
  - – 20 Credits: runs for the whole academic year.
  - – One exam [60%]:
    - In the summer.
    - Covering material from both semesters.
  - – Two submissions on the laboratories:
    - Second semester: ~ week 3 [25%].
    - Second semester: ~ week 11 [15%].

Timetabled Activities

- Workshops:
  - Structured input and exploratory activities.
  - No single style or balance: some group work, some trying of software, some discussions, some games.
  - Associated chapters or documents to provide a written record of the details.

- Laboratories:
  - Several timetabled sessions for trying out new systems or techniques (set tasks, known answers).
  - Several timetabled consultancy sessions for more open ended assessed project.
  - Submissions will involve both code and discussion of issues.

Organisation

- 1 Workshop: Introduction to unit (*this!*).
- 5 Workshops: Documents on the web.
- 3 Workshops: Services on the web.
- 4 Laboratories: Map Reduce for web document search.
- 5 Workshops: Data on the web.
- 4 Laboratories: Map Reduce for web data search.
- 1 Revision Session.

Position in Workshop

- Overview of unit.
- Documents on the web.
- Services on the web.
- Data on the web.

Documents on the Web

- What counts as a document?
- In *Information Retrieval*, a document is simply an object that can be
  - Indexed (so that it can be found again)
  - Searched for
- So, a document can be a text, an image, an audio file, a movie, …
- We'll concentrate here on documents as texts

Documents as texts

- "Unstructured data comprises the vast majority of data found in an organization. Some estimates run as high as 80%." (Merrill-Lynch, 1998)
- Are texts not typically structured?
  - Structured only for the human
  - Strictly unstructured for the machine

A few problems

- Documents (texts) are unstructured
- They are expressed in natural languages
- The information we seek may not be expressed in a language we know
- Natural language is highly ambiguous
- There is a vast and rapidly increasing number of documents
- People (and systems) have many and varied information needs

MANCHESTER
1824
The University of Manchester

Ambiguity

- FOOT HEADS ARMS BODY (The Times)
- Women bitten by rabid bat found in crates outside pub
- Mother and baby in pram hit by car (BBC)
- Time flies like an arrow. Fruit flies like a banana. (Oettinger)
- Police help dog bite victim
- I love my dog more than you
- Old men and women
- The soldiers shot the women and they fell down

Would you fly if this was in maintenance documentation?

- Remove bolt and stop (Barthe on need for ASD SE)

Some more problems

- Most documents are irrelevant for some query
  - How do we judge relevance?
- Most queries return large numbers of documents
  - How do we rank results?
- Most users are not information specialists
- Users are impatient and demand accuracy

Operating at Web scale

- What is involved in identifying and indexing at Web scale?
  - And in maintaining your index as the collection expands
- What should we index? How deep should/can we analyse?
- Trade off: depth vs. rapidity
  - Deep analysis takes time
- At Web scale, *any* analysis takes time
  - Offline analysis is the norm
  - Distributed, parallel computing is the norm

Object of retrieval?

- Commonly, what you get back is a ranked list of documents
  - Not an answer
  - You have to read the documents (or snippets)
    - They might or might not help you answer your information need
    - There may be some clever matching with lists of existing questions and answers, or clever use of logs and links
- Conclusion? Rather basic indexing techniques being used to support general search

The importance of order

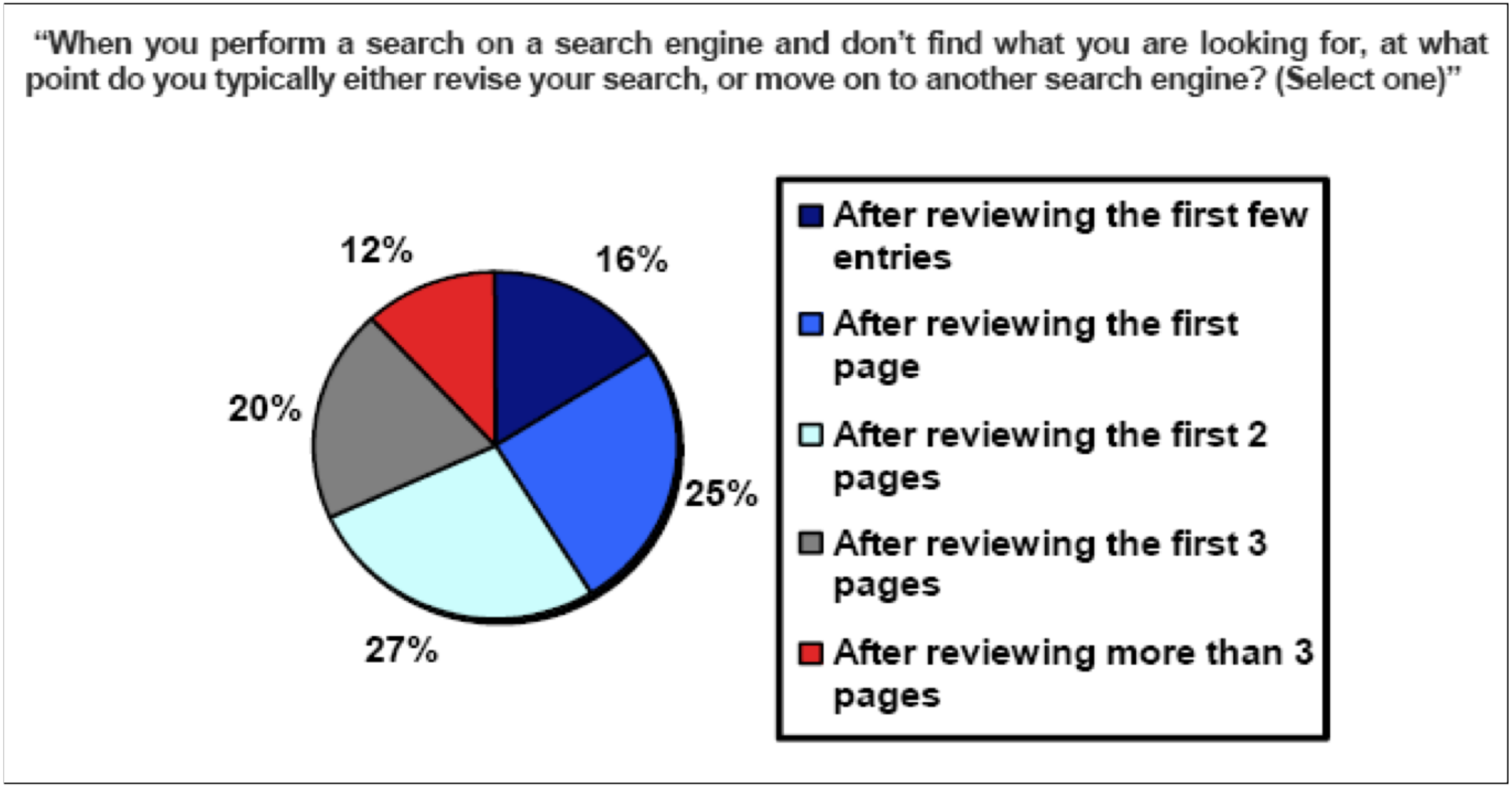
- Try a simple Google search for
  - blind venetian musicians
- Martino Pesenti is one such, but can you find him or information on any others on the first page?
- What does this tell you about indexing (and search)?

How big is the Web?

- See http://www.worldwidewebsize.com/
  - Search engine indexes overlap (duplicates)
  - So estimate size by subtracting estimated overlap
  - Difference pages/documents?
- But this is only the *indexed* Web
  - Small in comparison to the unindexed *deep* Web
  - Bergman, M.K. (2001) The Deep Web: Surfacing hidden value. *J. Electronic Publishing* 7(1)
  - DW 2001: estimated 550bn documents. Today (no recent figures…)?

MANCHESTER
1824
The University of Manchester

How far do people look for results?



"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"

(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Time we got our act together

“Bush urges that men of science should turn to the massive task of making more accessible our bewildering store of knowledge. Now, says Bush, instruments are at hand which, if properly developed, will give man access to and command over the inherited knowledge of the ages.”

When do you think this statement was made?

Are we there yet?

- “Google needs to move from words to meaning. [...] Google's long-term goal is to be able to give you one answer, which is exactly the right answer.” (Schmidt, Google, 2009)
- Google Hummingbird introduced in September 2013
  - Did you notice?
  - How is it different?
  - Is it better than what came before?

Resources

- http://www-nlp.stanford.edu/IR-book/
  - Chapter 1
- Liddy on Document Retrieval (book chapter)
  - http://surface.syr.edu/istpub/51/
- Bush's Memex
  - http://www.theatlantic.com/doc/194507/bush
- Video lectures (many technical, be selective)
- http://videolectures.net/Top/Computer_Science/Information_Retrieval/
- Das & Jain (Google) (book chapter)
  - http://cdn.dejanseo.com.au/wp-content/uploads/2012/04/Indexing-The-World-Wide-Web-The-Journey-So-Far.pdf

Position in Workshop

- Overview of unit.
- Documents on the web.
- Services on the web.
- Data on the web.

Cloud Services

- Service Provider Enterprises have grown to offer a wide range of services with high scalability, capable of processing large amounts of data over collections of commodity computers hosted by data centres.
- Because the costs of these service providers may be substantially lower than the costs of enterprises running their own data centres, several providers have made a business of “sharing” their data centres on a *pay-as-you-go* basis.
- The result is called public *cloud services*, which offer computing, storage and communication services, among others.

Typical Cloud Services Scenarios

- ***Scenario 1:*** A website records user activity for analysis of user behaviour, but cannot cope with the volume of data to be analysed. So, this enterprise links up with a Cloud service specialized in providing *processing power* in the form of a large clusters of computers able to execute the data analysis programs used by the website over large amounts of data, distributing the data across a cluster of computers and processing the data in in parallel.
- ***Scenario 2:*** A scientific laboratory is unable to store the large numbers of images obtained everyday from laboratory equipment and sensors, so it links up with a Cloud service that specialises in providing data storage services for other enterprises.

Cloud Service Providers

- Examples include: Amazon, Google, SAP, Oracle, Microsoft, Salesforce.com, etc.
- Amazon is among the largest cloud service providers, and offers a large number of services, such as:

Compute Capacity: Amazon Elastic Compute Cloud (EC2).	Database: Amazon Relational Database Service (RDS).
Massive Data Processing: Amazon Elastic MapReduce.	Payment and Billing: Amazon Flexible Payments Service (FPS).
Content Delivery: Amazon CloudFront.	Storage: Amazon Simple Storage Service (S3).

Activity

- Each desk should choose a cloud provider, and identify:
  - The name of their cloud offering.
  - The services supported.
  - The charging model.
  - Example customers.
- We will then review what you have discovered, and what this tells us about cloud services.

Example of a Cloud Service

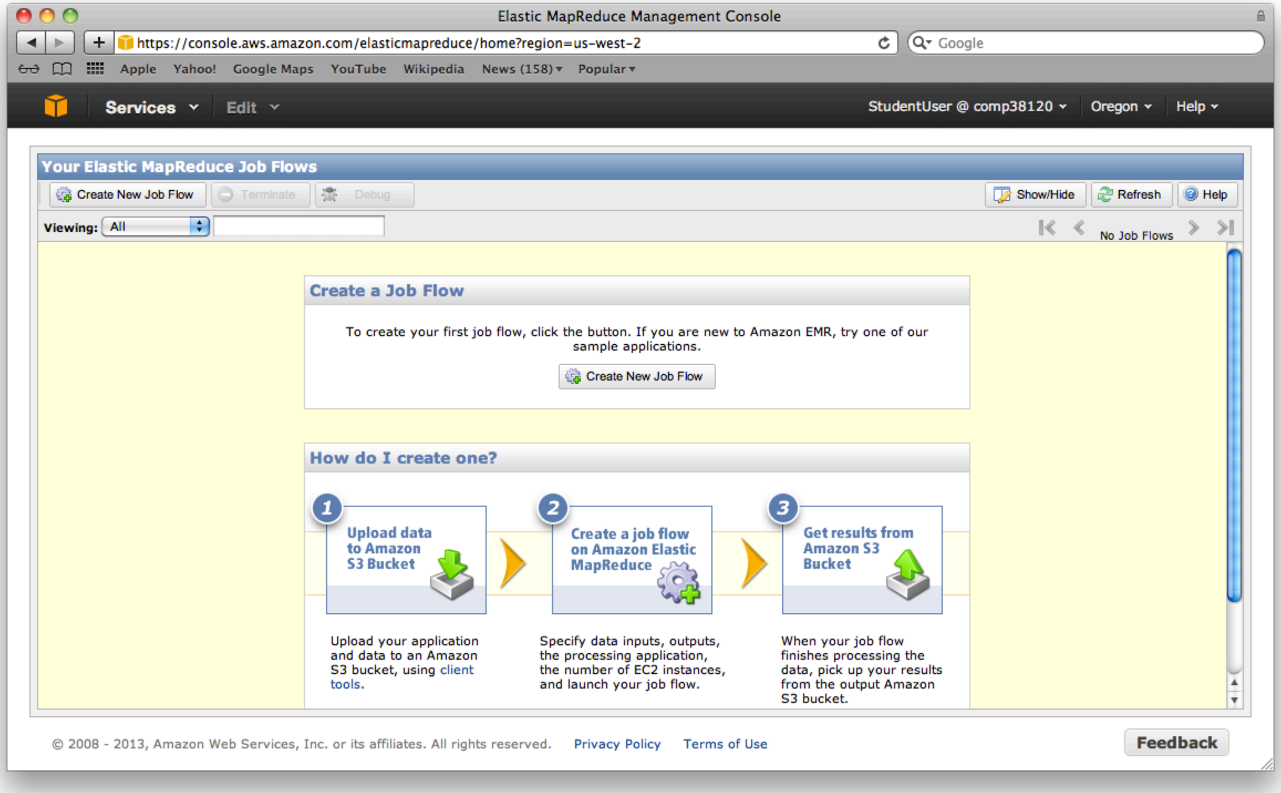
- One recurring feature of the web is scale; one of the features of cloud services is elasticity – the ability to change the amount of resource used to reflect changing needs.
- An activity associated with both documents and data on the web is search; web search involves ongoing crawling and index construction.
- Index creation from web crawls is often supported using the map-reduce paradigm.

Map Reduce (and the cloud) - 1

- *Map Reduce* is a scalable programming model, originally developed by Google for tasks such as index building.
- In Map Reduce:
  - applications are developed using two simple, functional operations (*map* and *reduce*) … and a few other supporting players;
  - the infrastructure supports the running of Map Reduce applications in parallel on potentially huge data sets on potentially numerous commodity machines.
- *Hadoop* is a widely used open source implementation of map reduce (hadoop.apache.org).

Map Reduce (and the cloud) - 2

- Amazon provides an Elastic Map Reduce (EMR) service that allows Hadoop programs to be run on their infrastructure (using other Amazon Web Services).



Create a Job Flow

To create your first job flow, click the button. If you are new to Amazon EMR, try one of our sample applications.

How do I create one?

1. Upload data to Amazon S3 Bucket
   Upload your application and data to an Amazon S3 bucket, using client tools.
2. Create a job flow on Amazon Elastic MapReduce
   Specify data inputs, outputs, the processing application, the number of EC2 instances, and launch your job flow.
3. Get results from Amazon S3 Bucket
   When your job flow finishes processing the data, pick up your results from the output Amazon S3 bucket.

© 2008 - 2013, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Later you will…

- Learn more about the different types of (mostly cloud) services on the web, their design, cloud economics, etc.
- Learn some approaches the economics of cloud computing – is it cost-effective to move to cloud?
- Learn more about application development using map reduce, including:
  - Hadoop – the infrastructure, its properties.
  - Map reduce – application design and patterns.
- Apply map reduce in labs to search applications in the web of documents and the web of data (Java, linux, hadoop, Eclipse).

Position in Workshop

- Overview of unit.
- Documents on the web.
- Services on the web.
- Data on the web.

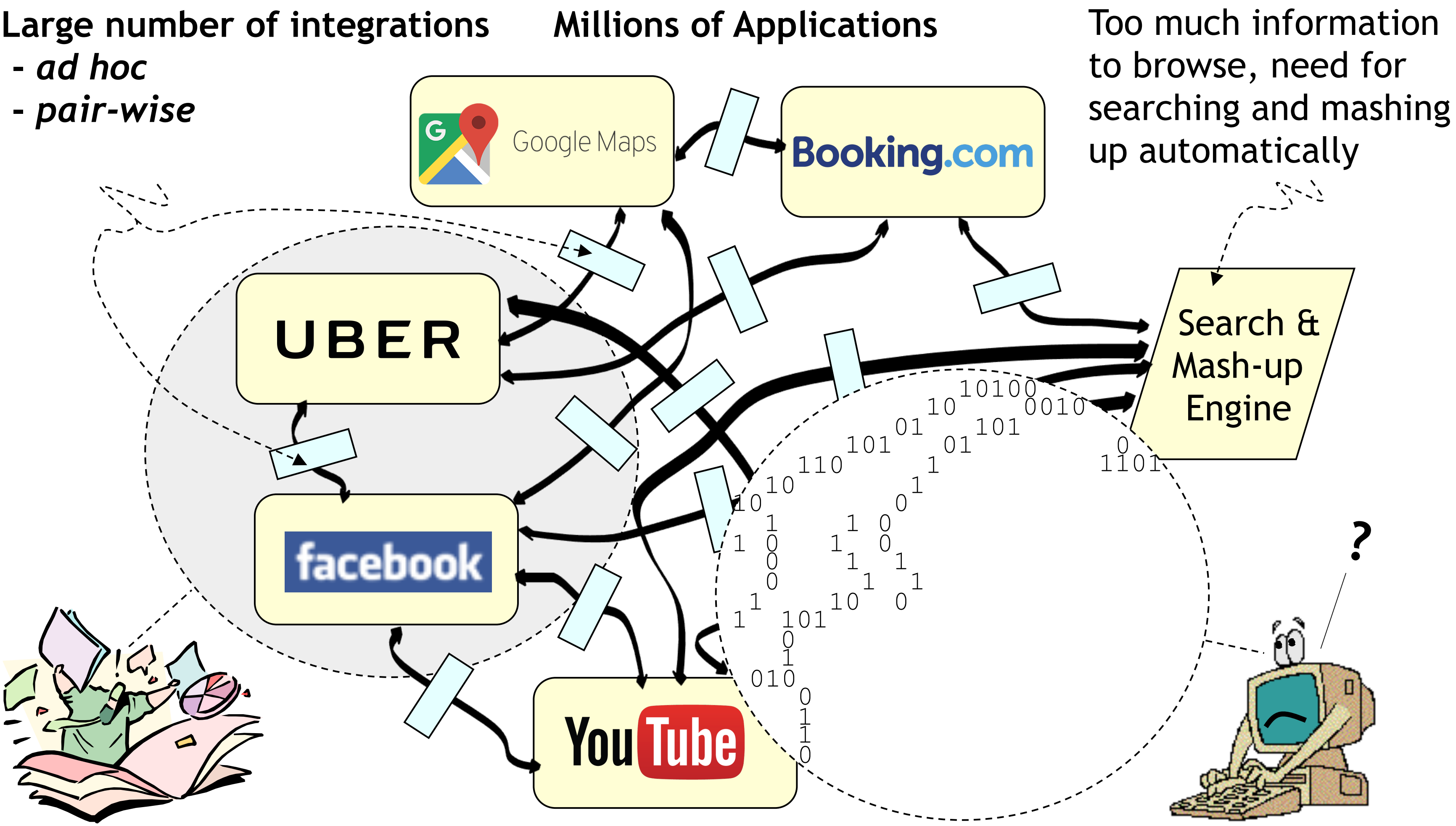
Position in Workshop

- Overview of unit.
- Documents on the web.
- Services on the web.
- Data on the web.

Web of Documents

- The Web is full of documents
  - – a global **file system.**
- Links between **documents** (or parts of them).
- Fairly **low** degree of structure
  - Mainly for visualisation and browsing.
- Designed for **human** consumption.

The Web Today



Large number of integrations
- *ad hoc*
- *pair-wise*

Millions of Applications

Too much information to browse, need for searching and mashing up automatically

Each site is “understandable” for us

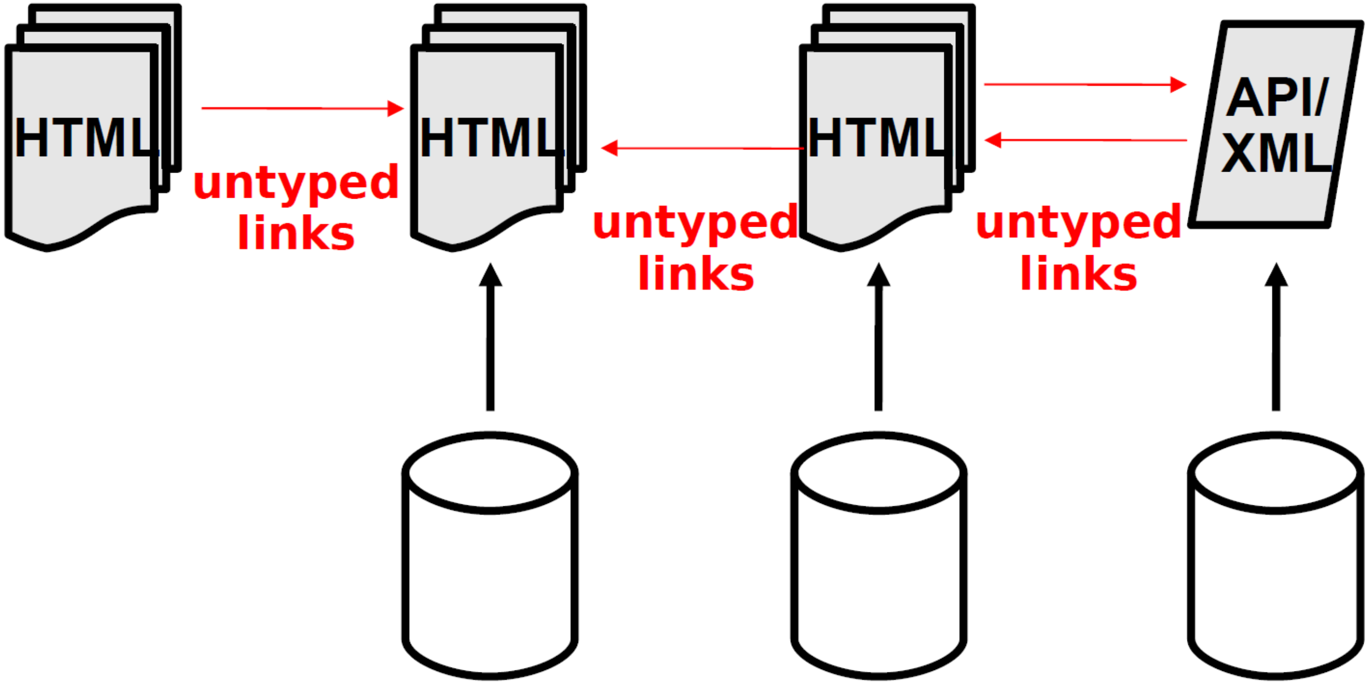
Computers don’t “understand” much

Web of Documents: Issues

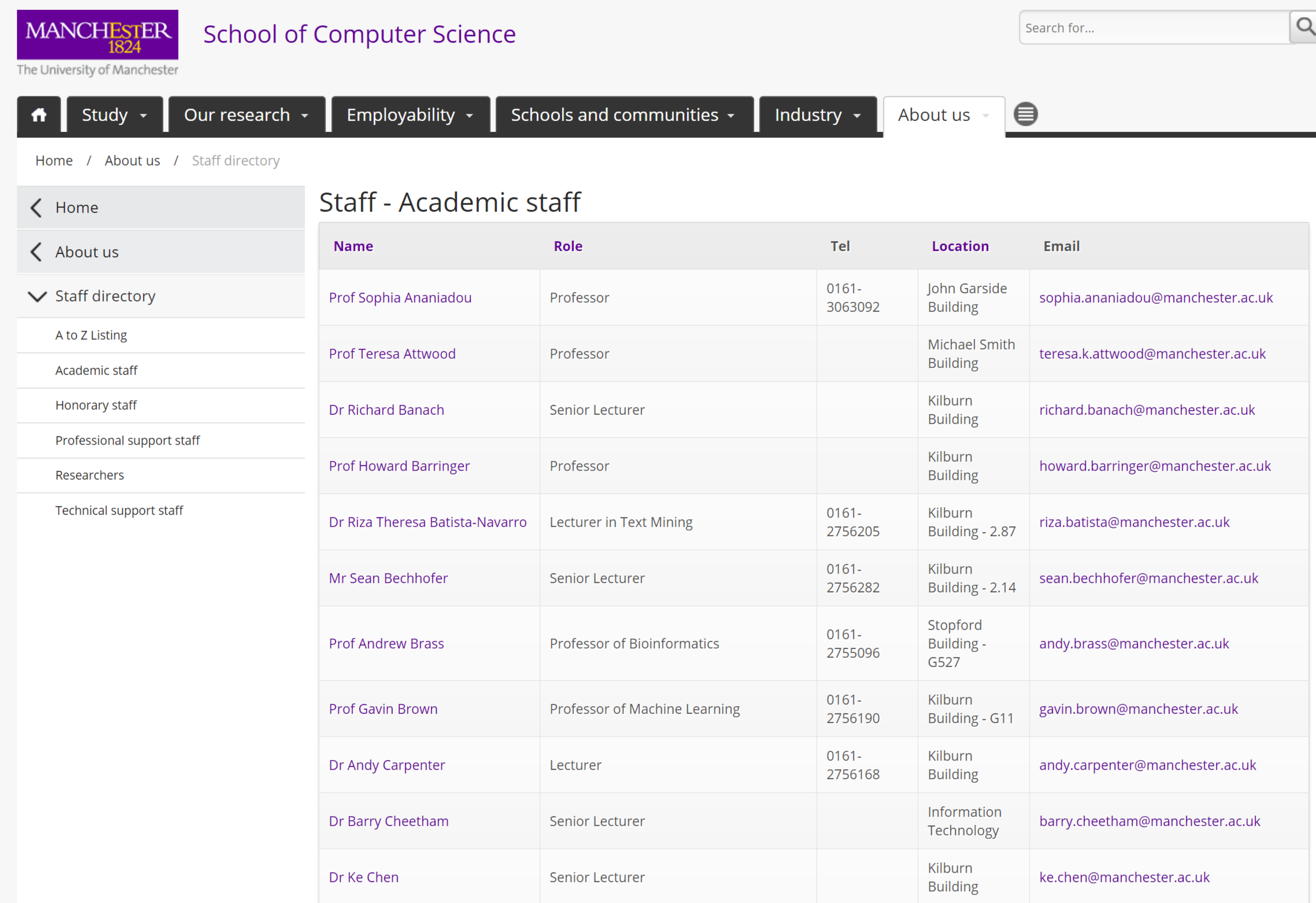
- One type of link ("untyped").
- We are rarely interested in documents, but rather in different *things:*
  - we need to read documents and *find* ourselves what we are *searching* for
  - difficult for machines because of **implicit semantics** of content and links.
- Disconnected data silos.

MANCHESTER
1824
The University of Manchester

Web of Documents

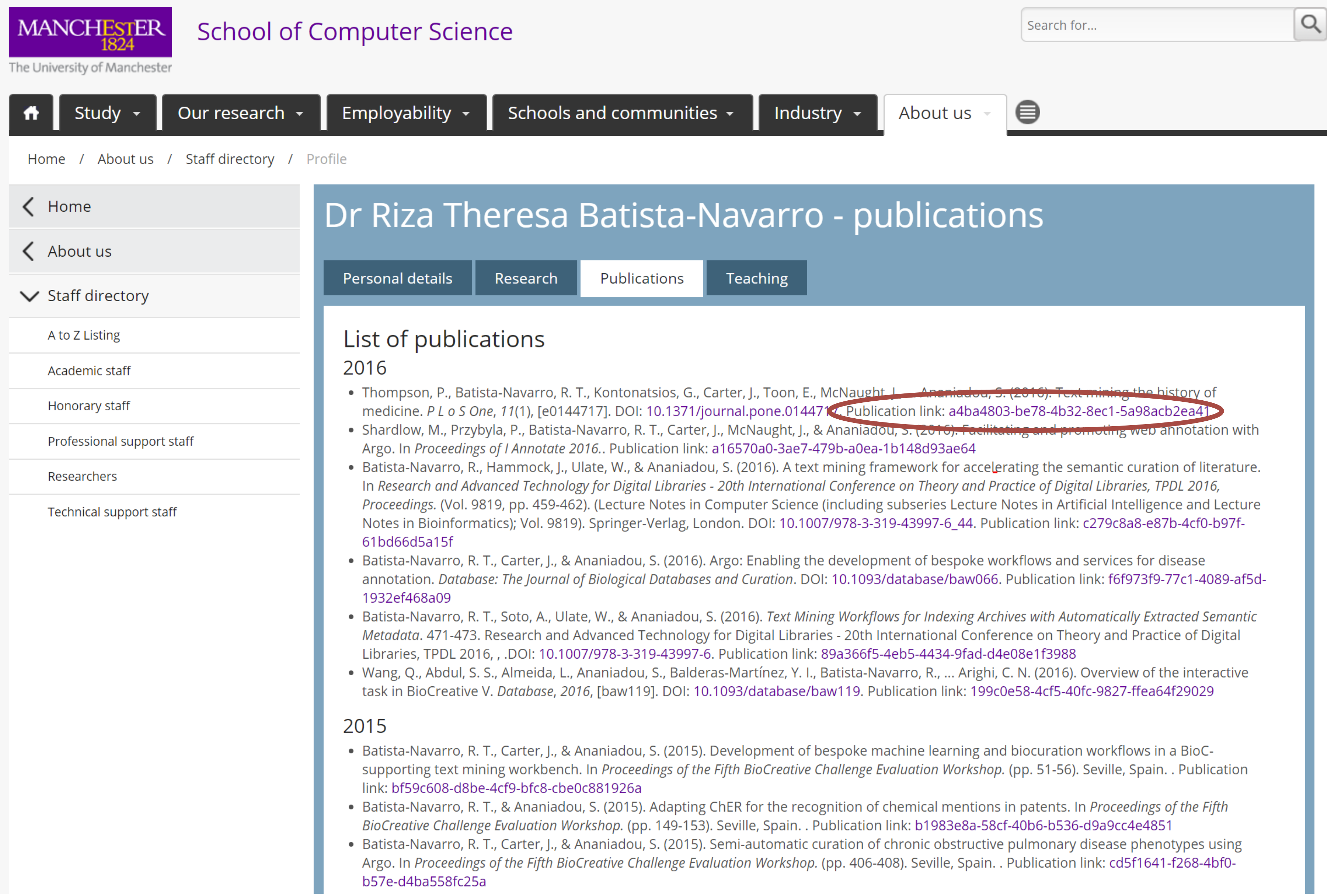


Taken from “An Introduction to Linked Data” by Tom Heath, 2009.
http://linkeddata.org/guides-and-tutorials



Staff - Academic staff

Name	Role	Tel	Location	Email
Prof Sophia Ananiadou	Professor	0161-3063092	John Garside Building	sophia.ananiadou@manchester.ac.uk
Prof Teresa Attwood	Professor		Michael Smith Building	teresa.k.attwood@manchester.ac.uk
Dr Richard Banach	Senior Lecturer		Kilburn Building	richard.banach@manchester.ac.uk
Prof Howard Barringer	Professor		Kilburn Building	howard.barringer@manchester.ac.uk
Dr Riza Theresa Batista-Navarro	Lecturer in Text Mining	0161-2756205	Kilburn Building - 2.87	riza.batista@manchester.ac.uk
Mr Sean Bechhofer	Senior Lecturer	0161-2756282	Kilburn Building - 2.14	sean.bechhofer@manchester.ac.uk
Prof Andrew Brass	Professor of Bioinformatics	0161-2755096	Stopford Building - G527	andy.brass@manchester.ac.uk
Prof Gavin Brown	Professor of Machine Learning	0161-2756190	Kilburn Building - G11	gavin.brown@manchester.ac.uk
Dr Andy Carpenter	Lecturer	0161-2756168	Kilburn Building	andy.carpenter@manchester.ac.uk
Dr Barry Cheetham	Senior Lecturer		Information Technology	barry.cheetham@manchester.ac.uk
Dr Ke Chen	Senior Lecturer		Kilburn Building	ke.chen@manchester.ac.uk



Dr Riza Theresa Batista-Navarro - publications

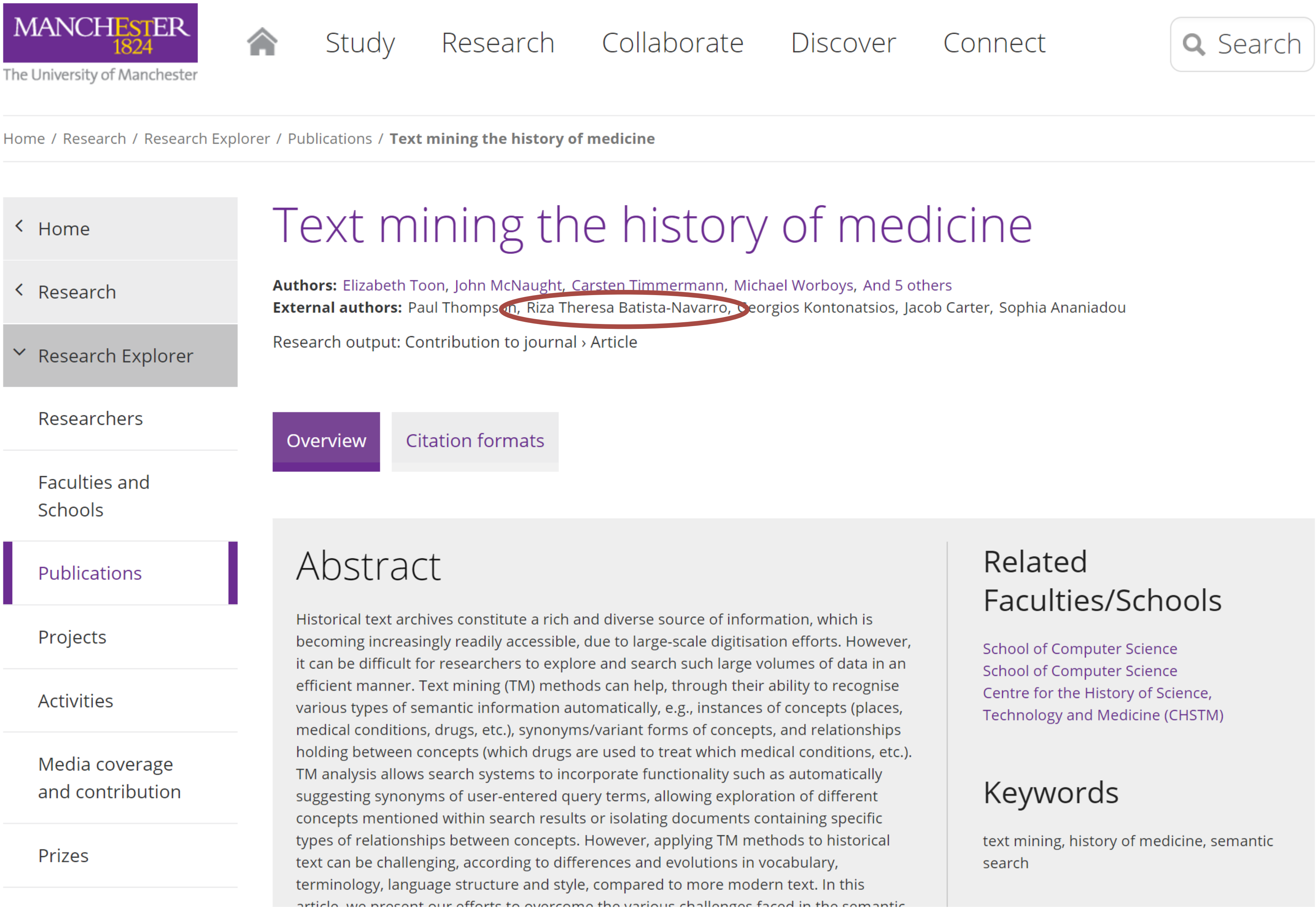
List of publications

2016

- Thompson, P., Batista-Navarro, R. T., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., ... Ananiadou, S. (2016). Text mining the history of medicine. *P L o S One*, *11*(1), [e0144717]. DOI: 10.1371/journal.pone.0144717. Publication link: a4ba4803-be78-4b32-8ec1-5a98acb2ea41
- Shardlow, M., Przybyla, P., Batista-Navarro, R. T., Carter, J., McNaught, J., & Ananiadou, S. (2016). Facilitating and promoting web annotation with Argo. In *Proceedings of I Annotate 2016.*. Publication link: a16570a0-3ae7-479b-a0ea-1b148d93ae64
- Batista-Navarro, R., Hammock, J., Ulate, W., & Ananiadou, S. (2016). A text mining framework for accelerating the semantic curation of literature. In *Research and Advanced Technology for Digital Libraries - 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Proceedings.* (Vol. 9819, pp. 459-462). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 9819). Springer-Verlag, London. DOI: 10.1007/978-3-319-43997-6_44. Publication link: c279c8a8-e87b-4cf0-b97f-61bd66d5a15f
- Batista-Navarro, R. T., Carter, J., & Ananiadou, S. (2016). Argo: Enabling the development of bespoke workflows and services for disease annotation. *Database: The Journal of Biological Databases and Curation*. DOI: 10.1093/database/baw066. Publication link: f6f973f9-77c1-4089-af5d-1932ef468a09
- Batista-Navarro, R. T., Soto, A., Ulate, W., & Ananiadou, S. (2016). *Text Mining Workflows for Indexing Archives with Automatically Extracted Semantic Metadata*. 471-473. Research and Advanced Technology for Digital Libraries - 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, , .DOI: 10.1007/978-3-319-43997-6. Publication link: 89a366f5-4eb5-4434-9fad-d4e08e1f3988
- Wang, Q., Abdul, S. S., Almeida, L., Ananiadou, S., Balderas-Martínez, Y. I., Batista-Navarro, R., ... Arighi, C. N. (2016). Overview of the interactive task in BioCreative V. *Database*, *2016*, [baw119]. DOI: 10.1093/database/baw119. Publication link: 199c0e58-4cf5-40fc-9827-ffea64f29029

2015

- Batista-Navarro, R. T., Carter, J., & Ananiadou, S. (2015). Development of bespoke machine learning and biocuration workflows in a BioC-supporting text mining workbench. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop.* (pp. 51-56). Seville, Spain. . Publication link: bf59c608-d8be-4cf9-bfc8-cbe0c881926a
- Batista-Navarro, R. T., & Ananiadou, S. (2015). Adapting ChER for the recognition of chemical mentions in patents. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop.* (pp. 149-153). Seville, Spain. . Publication link: b1983e8a-58cf-40b6-b536-d9a9cc4e4851
- Batista-Navarro, R. T., Carter, J., & Ananiadou, S. (2015). Semi-automatic curation of chronic obstructive pulmonary disease phenotypes using Argo. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop.* (pp. 406-408). Seville, Spain. . Publication link: cd5f1641-f268-4bf0-b57e-d4ba558fc25a



Text mining the history of medicine

Authors: Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys, And 5 others
External authors: Paul Thompson, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Sophia Ananiadou

Research output: Contribution to journal › Article

Abstract

Historical text archives constitute a rich and diverse source of information, which is becoming increasingly readily accessible, due to large-scale digitisation efforts. However, it can be difficult for researchers to explore and search such large volumes of data in an efficient manner. Text mining (TM) methods can help, through their ability to recognise various types of semantic information automatically, e.g., instances of concepts (places, medical conditions, drugs, etc.), synonyms/variant forms of concepts, and relationships holding between concepts (which drugs are used to treat which medical conditions, etc.). TM analysis allows search systems to incorporate functionality such as automatically suggesting synonyms of user-entered query terms, allowing exploration of different concepts mentioned within search results or isolating documents containing specific types of relationships between concepts. However, applying TM methods to historical text can be challenging, according to differences and evolutions in vocabulary, terminology, language structure and style, compared to more modern text. In this article, we present our efforts to overcome the various challenges faced in the semantic

Related Faculties/Schools

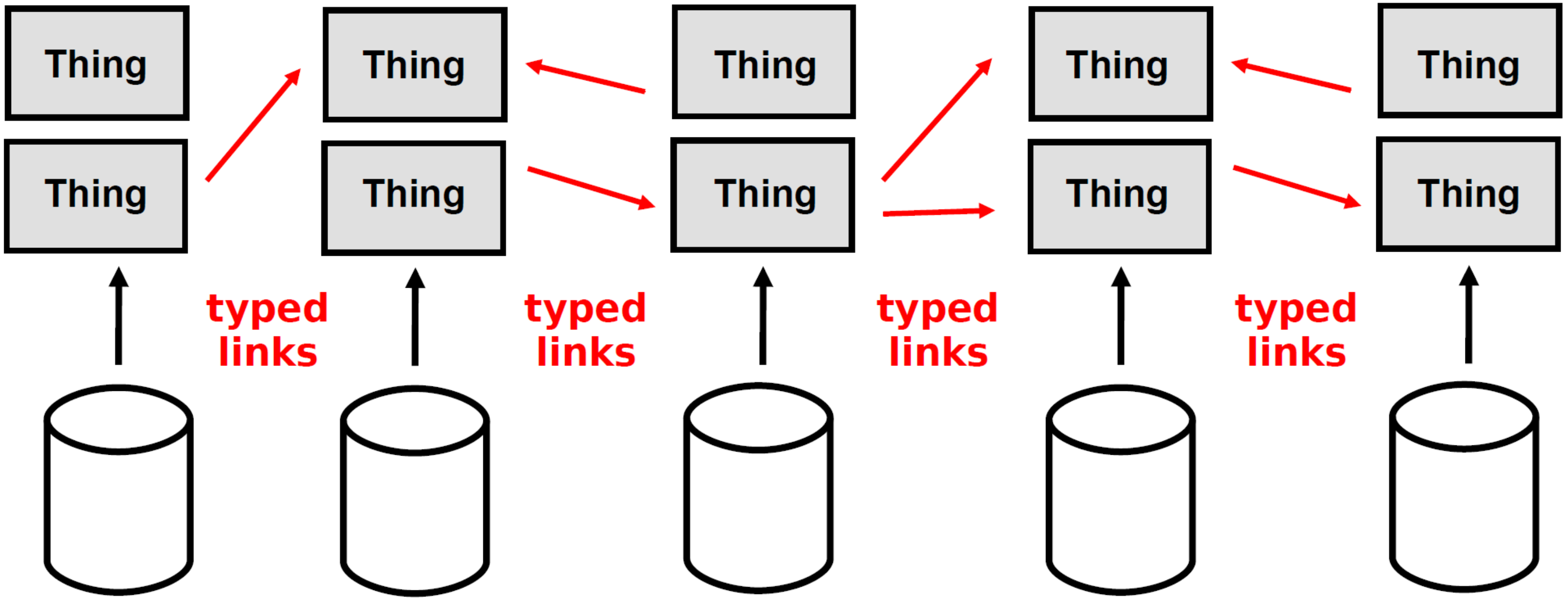
School of Computer Science
School of Computer Science
Centre for the History of Science, Technology and Medicine (CHSTM)

Keywords

text mining, history of medicine, semantic search

MANCHESTER
1824
The University of Manchester

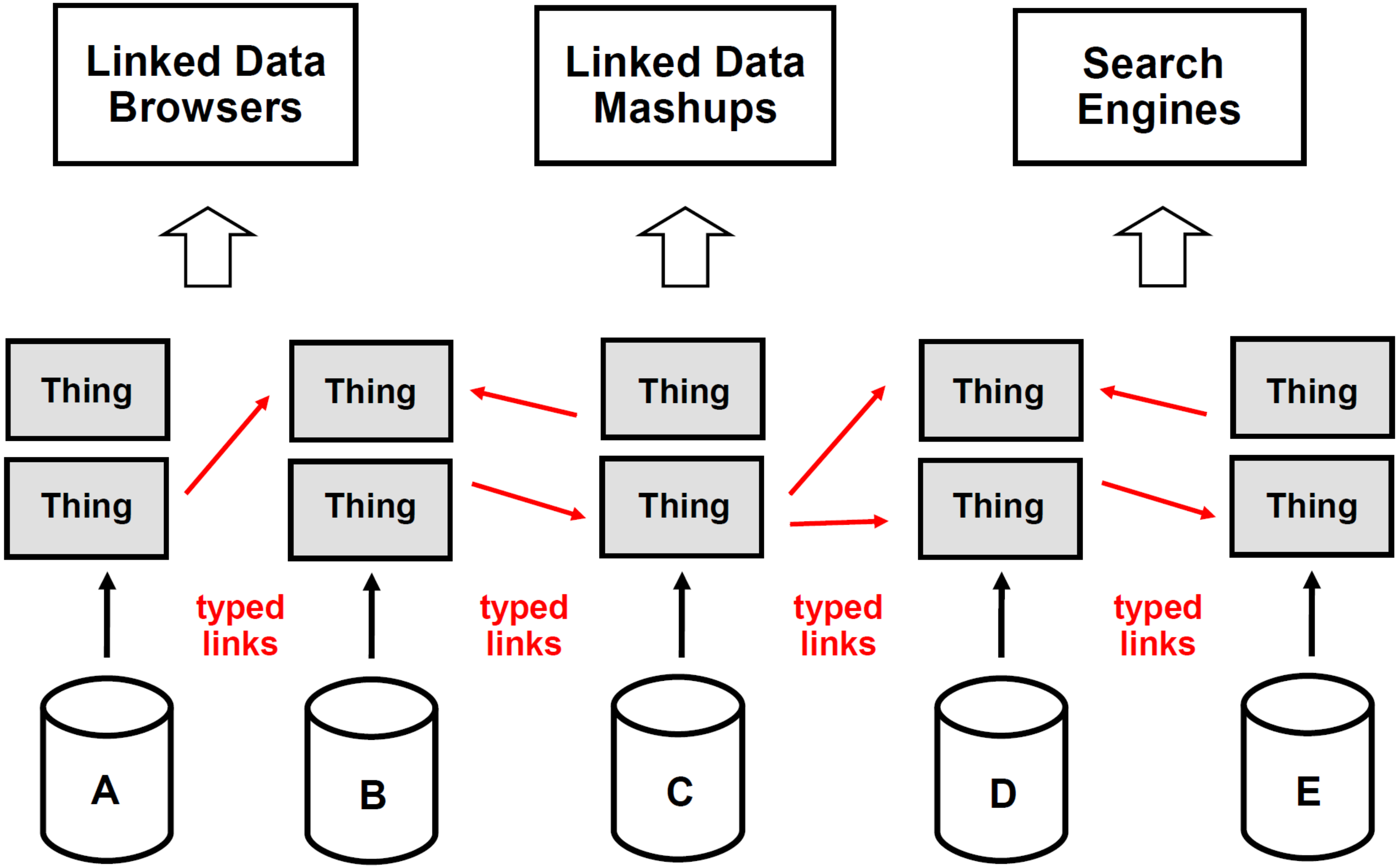
Web of Data



Taken from “An Introduction to Linked Data” by Tom Heath, 2009.
http://linkeddata.org/guides-and-tutorials

MANCHESTER 1824
The University of Manchester

Web of Things



Taken from "An Introduction to Linked Data" by Tom Heath, 2009.
http://linkeddata.org/guides-and-tutorials

Web of Data

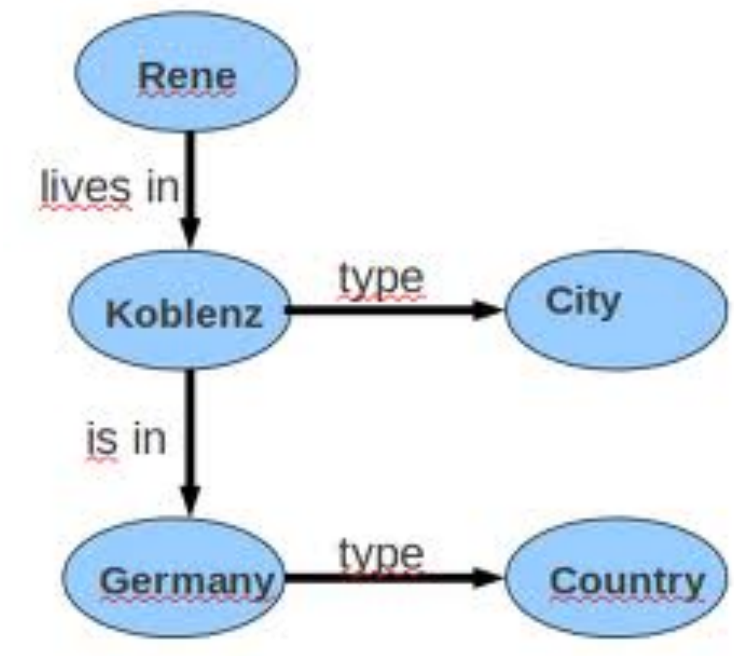
- Publish data on the Web, so that machines can understand it
  - have data in a global **database.**
- Links between **things** (i.e. description of things through data).
- Add **explicit** semantics.
- Designed for **machines** first so that data can be easily integrated.

Web of Data: Issues

- Two main issues:
  - *Different types of data* (CSV, XML, HTML, JSON, relational tables): can we find a **common format** to package the data (like we have for documents)?
  - *Linking*: how do we link these datasets together? Do we need **unique IDs** (like we have for documents)?

Web of Data: Format

- Subject – Predicate – Object triples



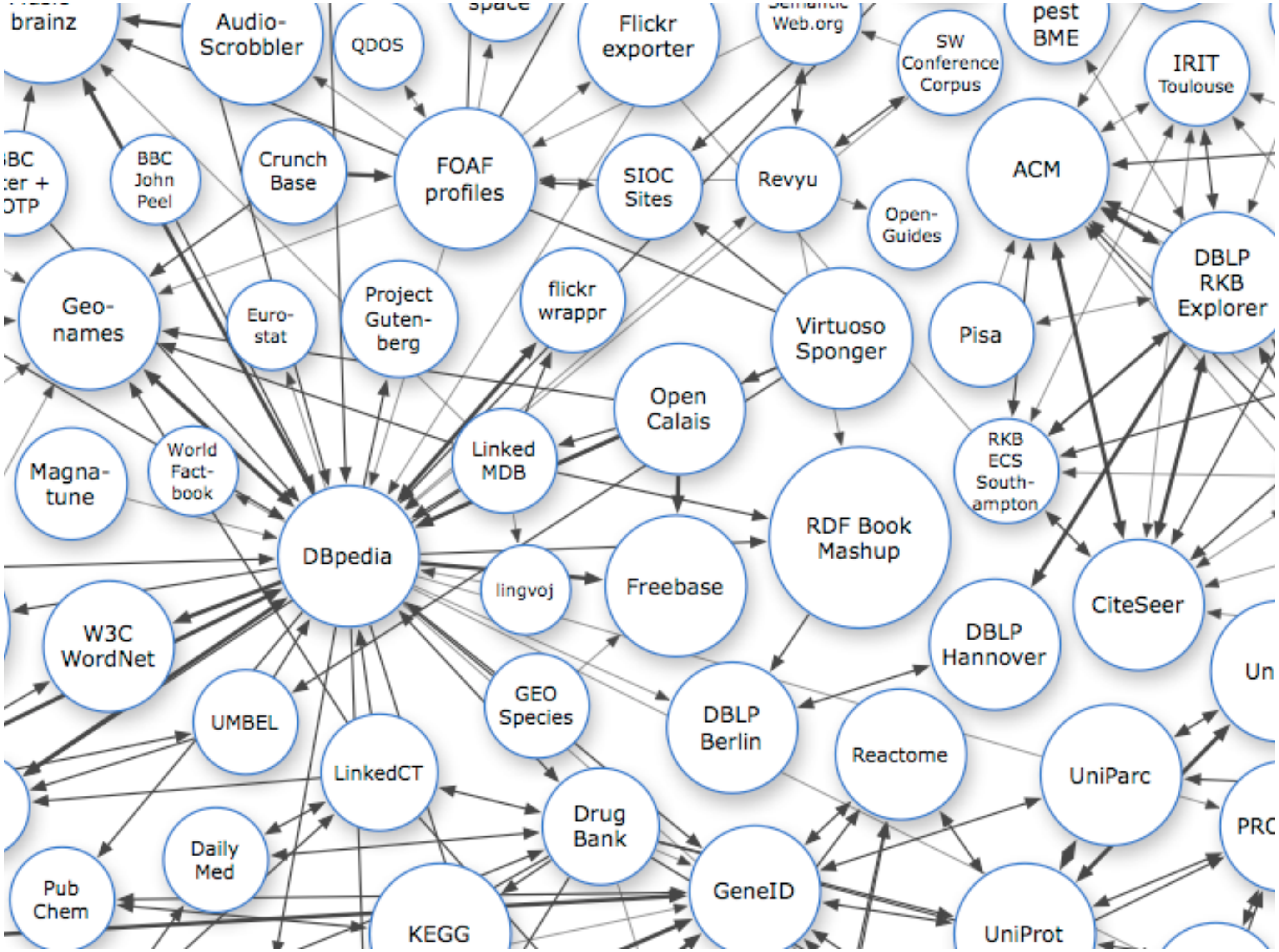
RDF triples

Web of Data: IDs

- Use Uniform Resource Identifiers (URI)
  - – analogous to URLs
  - – used for subjects, objects, predicates!

Hello, I am http://dbpedia.org/resource/
Homer_Simpson

Web of Data: Linking Linked Data



Web of data: DBpedia

- DBpedia - a linked data 'version' of Wikipedia.
- Visit http://dbpedia.org/fct/
  - select Entity Label Lookup
  - search for
    - Manchester and your home country
    - Homer Simpson
- Compare to Web of documents search

  (e.g. through Google).

Web of data: Wikidata

- http://www.wikidata.org/ - "a free, linked database that can be read and edited by both humans and machines".
- Search for
    - Manchester and your home country
    - Homer Simpson
    - Manchester United
- Compare to Web of documents search (e.g. through Google).

Discussion topics

- What kind of information did you find?
- Was it integrated?
- Quality of data?
- What could we do with this kind of data?
- How would you query this data?

Web of Data in COMP38120

In this module we will learn about:

- Principles of linked data.
- Linked data design.
- Publishing linked data.
- Consuming and aggregating linked data.
- Case study: using linked data in a real-world setting (BBC)