The University Of Sheffield.

Semantic Web Technologies for Knowledge Management

Prof. Fabio Ciravegna
Organisations Information and Knowledge
Department of Computer Science
University of Sheffield
United Kingdom

© Fabio Ciravegna, University of Sheffield

Organisations, Information and Knowledge

The University Of Sheffield.

Copyright Notice

- These slides were presented at The Seventh Summer School on Ontological Engineering and the Semantic Web (SSSW'09), July 5, 2009 - July 11, 2009. Cercedilla, near Madrid (Spain)
- They have been developed by Fabio Ciravegna,fabio@dcs.shef.ac.uk who keeps all the copyright of the material.
- You are very welcome to reuse these slides under the following conditions:
  - the use is limited to personal or educational purposes
  - the copyright note must always be visible when slides are presented
  - full recognition is given to me for the paternity of the slides and of the information contained
  - the context in which the slides are used/presented must be appropriate and not damaging of the image of the University of Sheffield or mine.
- Also I would be very grateful if you could let me know that you have reused them. Knowing that the result of your work is useful to others is always nice reassuring knowledge.

© Fabio Ciravegna, University of Sheffield

Organisations, Information and Knowledge

The University Of Sheffield.

Outline

- What is KM and what are its requirements
  - Requirements for Semantic KM
- Knowledge Acquisition in KM
  - Ontologies for KM
- Knowledge Capture in KM
  - Annotation of resources
- Knowledge Retrieval
- Knowledge Sharing and Reuse
  - Visualisation
- Conclusions

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Knowledge Management

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

From Data to Knowledge

- *Data*
  - the product of observation,
  - discrete and factual,
  - lacks structure, interpretation and meaning;
- *Information*
  - a transformation of data into a more effective and usable form,
  - via the process of structuring and organising;
- *Knowledge*
  - a collection of data, information, personal notes and reflections, related and coherent, purposefully collected with the intent to foster understanding of a specific phenomenon
  - "might be viewed as a mix of information, understanding, capabilities, experience, skills and values"

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Why Manage Knowledge?

- To enable easy timely and effective reuse of knowledge
  - We need: to enable retrieval
    - Requirements: easy and effective retrieval
    - Information/Data compatibility
- To enable retrieval
  - we need to: acquire knowledge sharing
- To enable sharing
  - we need to: acquire knowledge
    - Desiderata:
      - Easy acquisition (do not get in the way of the user's work!)
      - Comprehensive acquisition (do not miss important facts!)
- To enable acquisition:
  - We need modelling the domain and process in an appropriate way

© Fabio Ciravegna, University of Sheffield

What we know and what we do not

- As we know, there are known knowns
  - That are things we know we know.
- We also know there are known unknowns;
  - That is to say we know there are some things we do not know.
- But there are also unknown unknowns
  - The ones we don't know we don't know

Donald Rumsfeld

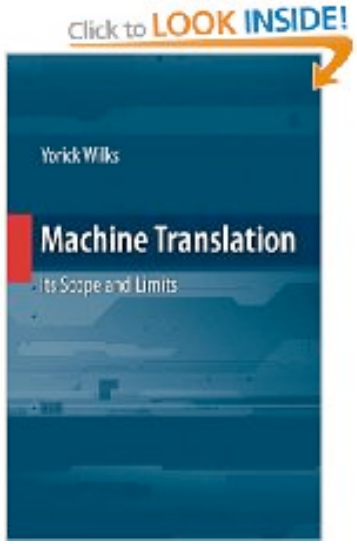
© Fabio Ciravegna, University of Sheffield

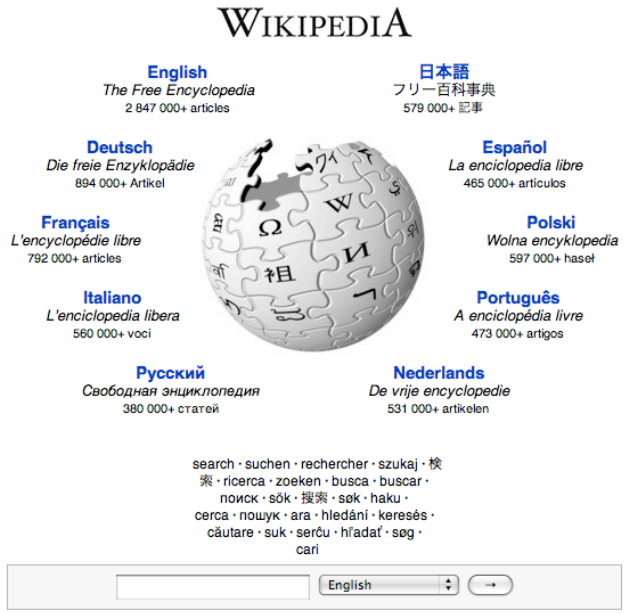
The University Of Sheffield.

Steps in KM

Share your known knowns

So that others can put them to use





http://www.wikipedia.org/

© Fabio Ciravegna, University of Sh

Steps (ctd)

known knowns

Share your known unknowns

Because they can be known knowns to others

i.e. Ask around if anyone knows the answer

© Fabio Ciravegna, University of Sheffield

Steps (ctd)

known unknowns

Google UK

Investigate your known unknowns

To turn them into known knowns

© Fabio Ciravegna, University of Sheffield

Steps (ctd)

Identify your
unknown unknowns

To turn them into known unknowns

known unknowns

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Trends in KM

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Trends: outline

- Services Vs Products
- Absolute Vs Local knowledge
- Outsourcing
- Large scale information space
- Multi and cross-mediality
- New knowledge Vs legacy data

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Product centred life-cycle

- Selling of complete service solutions instead of physical goods
  - aircraft power Vs jet engines
  - 7 year warranty on cars
- Requires
  - Taking charge of the whole product life-cycle
    - Designing better products to have larger margins
      - As opposed to design to manufacture at low cost
    - Design products to minimise service requirements
      - As opposed to profit on service provision

▪ 75% of Jet engine company's revenues from service

▪ Smaller and more efficient service facilities

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Product life-cycle & KM

- Knowledge organisation centred around Products and Services
  - As opposed to traditional design/manufacturing/service department
- Company-wide availability of knowledge
  - Beyond traditional boundaries
- Knowledge created across boundaries
  - E.g. by working groups or communities of practice

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Dynamic communities

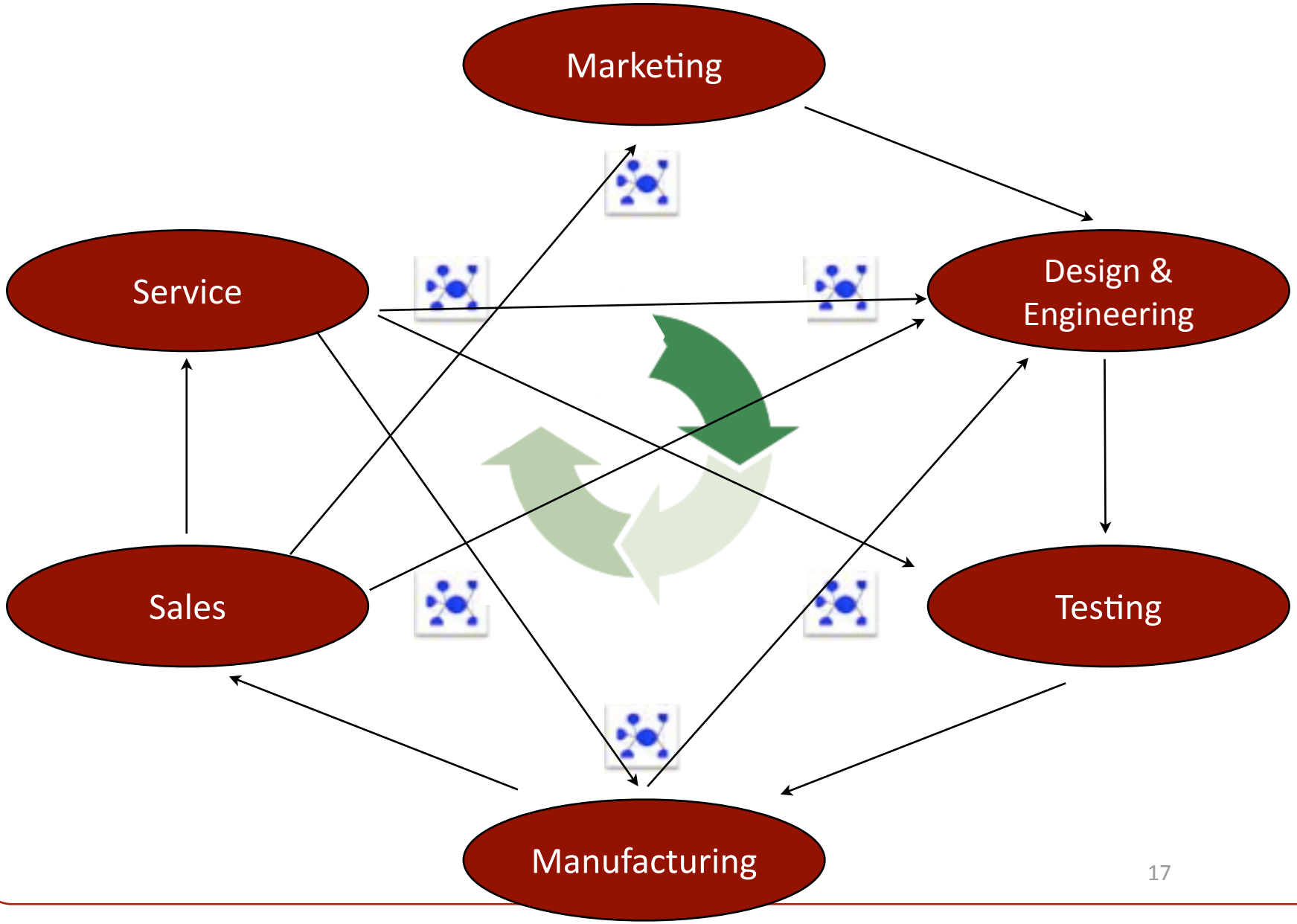
- Organisations based on dynamic communities
  - that acquire and share knowledge according to dedicated schemas
  - existing across traditional organisational boundaries
    - ill fit pre-determined standard schemas
    - require rapidly tailoring knowledge for their specific ad-hoc uses
    - often outside the company (outsourcing)

- Requirement:
  - independence of ontological views
  - communities must share their knowledge with rest of organisation

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

A product's life-cycle



© Fabio Ciravegna, University of Sheffield

Organisations, Information and Knowledge

The University Of Sheffield.

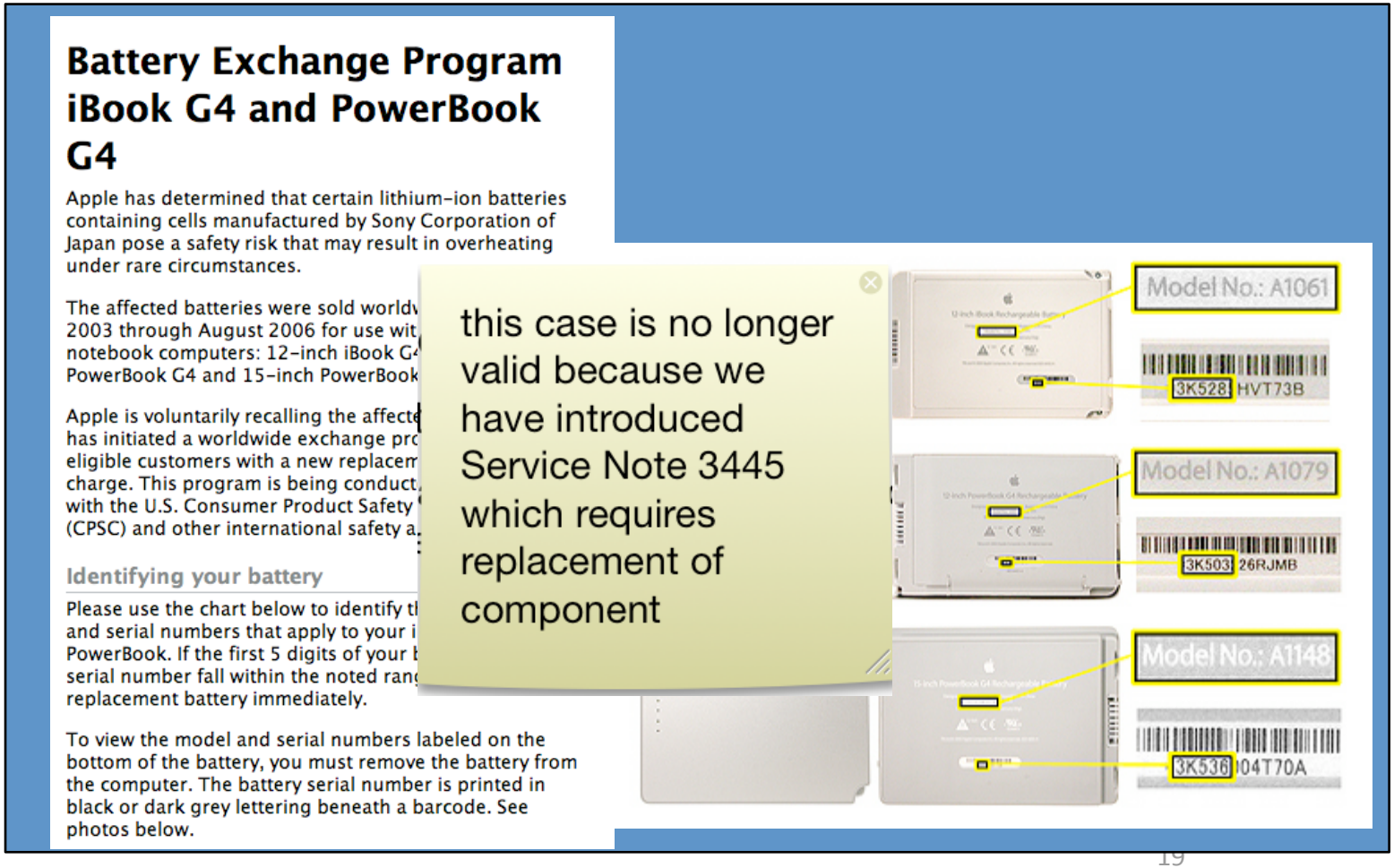
Outsourcing

- Outsourcing mainly used for manufacturing in the past
  - But increasingly for knowledge intensive tasks such as design or research
    - Aerospace and pharmaceutical companies
- Implication on knowledge management
  - Developed knowledge must be fed back into the company

© Fabio Ciravegna, University of Sheffield

Requirements for KA: Cross Media

- Evidence is often distributed in different media;
- Knowledge in one medium does not carry the full evidence

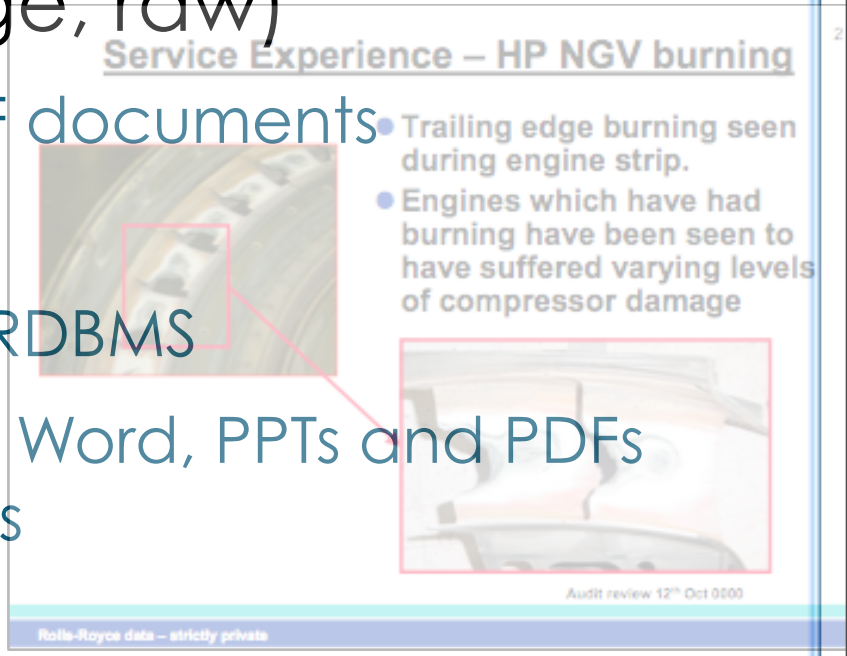


© Fabio Ciravegna, University of Sheffield

Compound Documents & CM

From Deliverable D8.2

- Typical data objects (text, image, raw)
  - Text formats: Word, Excel, PPT and PDF documents
  - Images: Jpeg and Gif
  - Raw data: Measurements stored in a RDBMS
  - Cross-media: Compound documents: Word, PPTs and PDFs containing both text and Jpeg images
    - Portions semantically related to each other
    - Information contained in just one modality is insufficient
    - Cross-media knowledge acquisition techniques needed to capture and manage explicit and implicit knowledge



© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Knowledge Management 2010+

- Knowledge management is shifting towards large scale
  - Dramatic reduction of memory and storage cost
  - Increase in speed and capacity
- Number and size of distributed archives of information
  - Large organisations' intranets as mini webs
    - Thousands of computers
    - Dozens of repositories
    - Dozens of millions of documents
      - soon to become hundreds of millions of documents
        - » size comparable to the Internet at the end of the 90s.
- Massive shift towards multi-media

© Fabio Ciravegna, University of Sheffield

X-MEDIA

Information Society Technologies

- Large Scale Knowledge Management across Media
- 2nd largest project funded on Kno Content Technologies by the EU:
  - €13.6M budget
- 2006-2010
- 15 partners
  - Users: Fiat, Rolls Royce
- In industrial board:
  - Kodak, Philips, Telenor, CESI, DS&S
- Project Director:
  - Prof. Fabio Ciravegna (U. Sheffield)

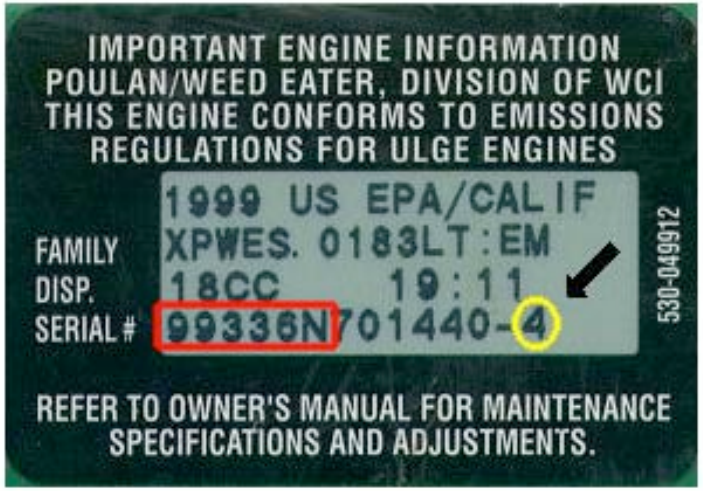
© Fabio Ciravegna, University of Sheffield

Organisations, Information and Knowledge

An Example of Knowledge Management

jet engines are moving towards complete serialisation
- every piece has a serial number (excepts nuts and bolts)
- the history of each part is recorded
  - e.g. part transferred between engines

© Rolls-Royce plc



99336N = Date Code
- Day of the Year
- Year of Production

4 = Product Type

© Fabio Ciravegna, University of Sheffield

Sense Making in Aerospace

- a jet engine can produce ~1Gbyte of vibration data per hour of flight;
  - if irregularities are found, part of the data can be stored
  - reports can be written (event reports)
  - pictures can be taken

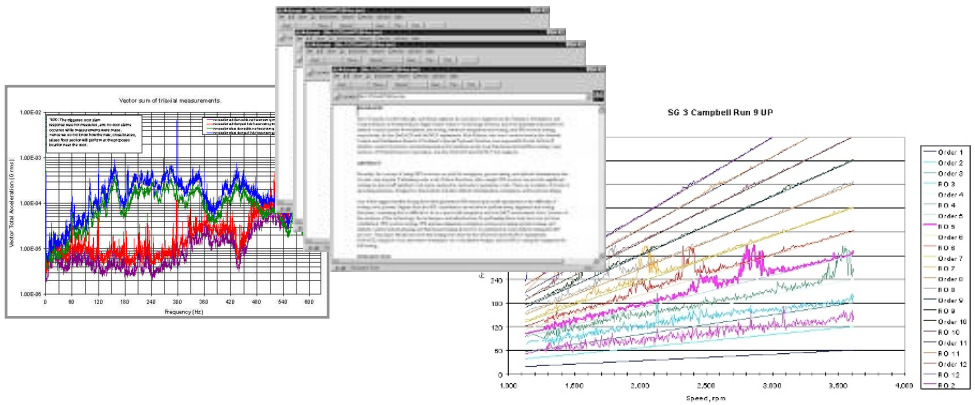


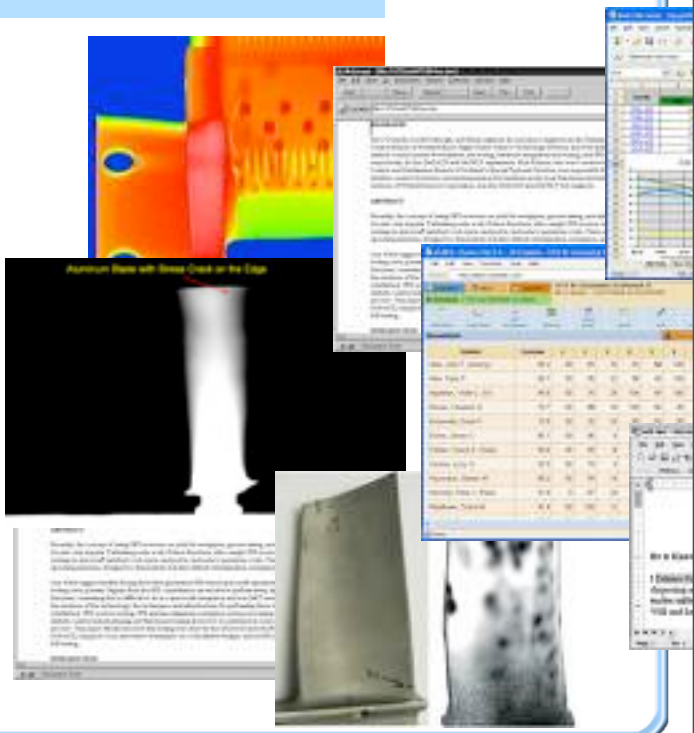
image © www.rolls-royce.com

© Fabio Ciravegna, University of Sheffield

Jet engine example (3)

When engine is serviced (e.g. overhaul)

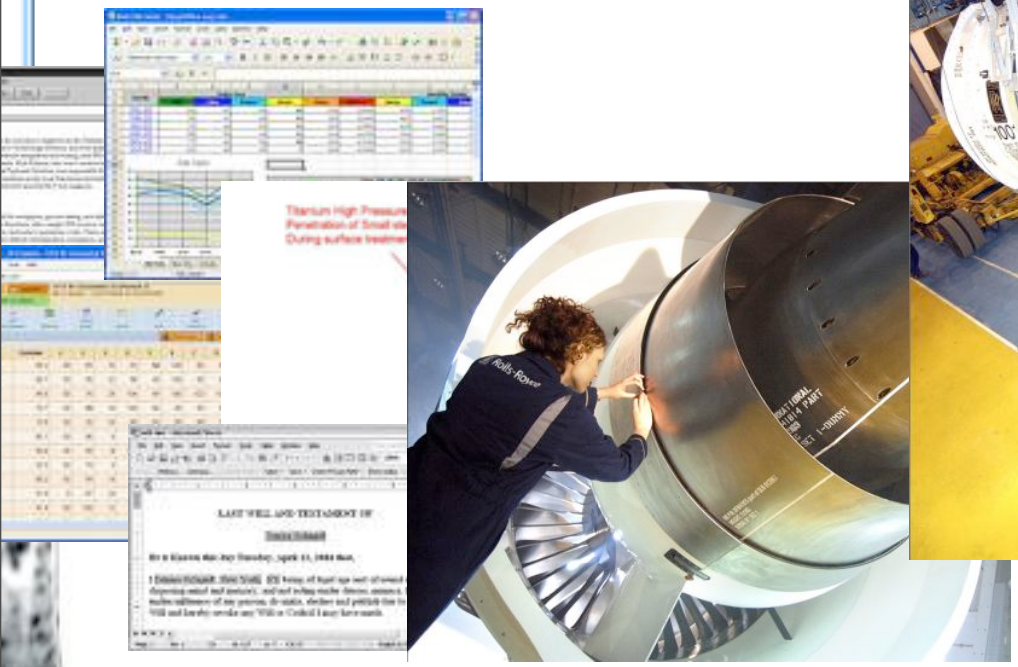
- financial information is produced.
- if issues are found,
  - pictures are taken
  - reports are written
  - engine is tested



© Fabio Ciravegna, University of Sheffield

Jet engine example (4)

- If problem is recurring (or suspected so)
  - a problem resolution group is established
    - existing evidence is retrieved
    - further evidence is collected
    - a learned lesson is generated
    - same problem is investigated across models



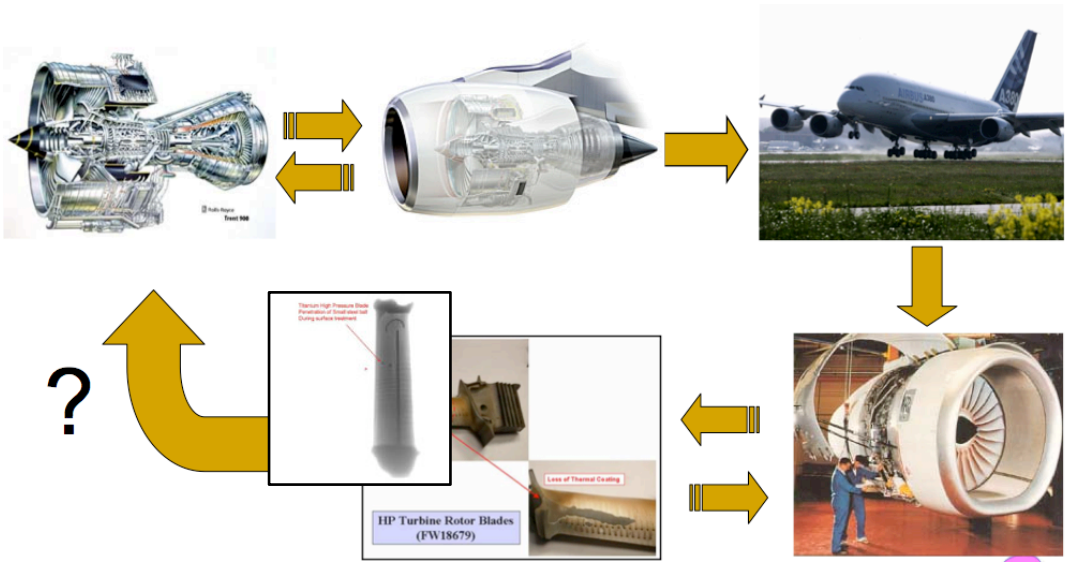
images © www.rolls-royce.com

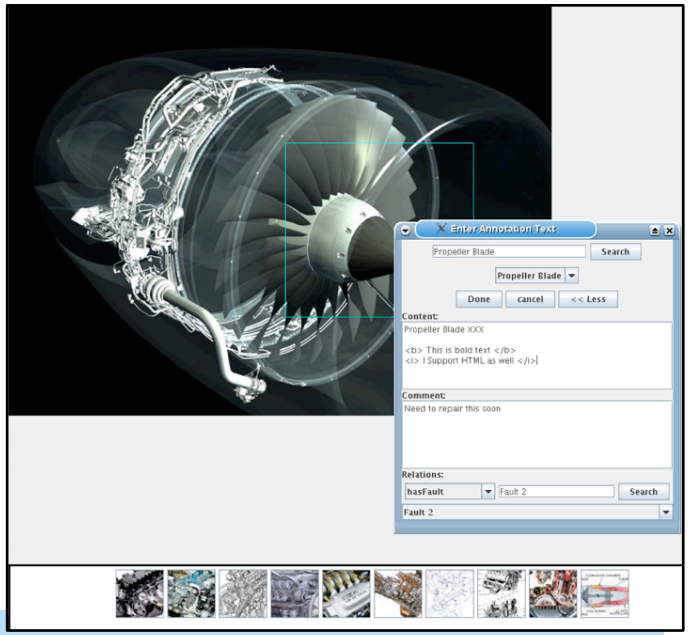
Document Type
AROC proforma
AROC results
Development
EHM data
Emails
ONWING emails
Images
Lab findings
Monitoring Requir
Presentations
Procedures
RCP
Risk Assessment
Solution Reports
Technical Reports
TS&O Reports

© Fabio Ciravegna, University of Sheffield

Jet engine example (4)

- Lifecycle "folder" will easily sum up to several Terabytes
- Folder will contain highly interrelated information stored in different media





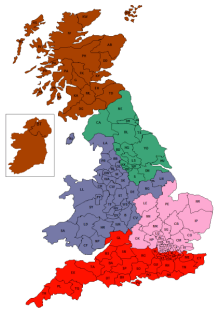
- Goal for Knowledge Management:
- Making information available independently from
  - Data format (structured/unstructured)
  - The archive
- Making it available for automatic processing
- Making it easily accessible and manageable despite its size

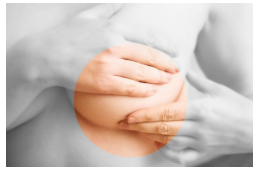
© Fabio Ciravegna, University of Sheffield

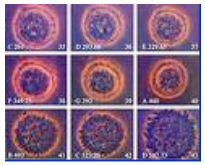
The University Of Sheffield.

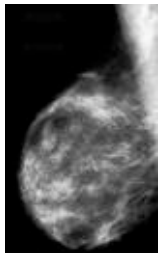
Other applications

- Tenders at Royal Mail
  - 15 people working full time for 2 weeks on tender offers
- Oil drilling
  - Rig debugging
  - Drilling conditions
- Triple assessment for breast canc
  - Radiologist (mammogram), breast surgeon and pathologist
- Fiat: competitor analysis









© Fabio Ciravegna, University of Sheffield

Organisations, Information and Knowledge

Image source: Google Images

The University Of Sheffield.

USfd Advance Manufacturing Centre

The University Of Sheffield.

BAE Systems is a global company engaged in the development, delivery and support of advanced defence and aerospace systems in the air, on land and at sea.	Boeing is the world's leadingaerospace company and the largest manufacturer of commercial jetliners and military aircraft combined.	CG Tech specializes in numerical control simulation, verification, optimization, and analysis software technology for manufacturing.
STUART D A Stuart is recognized as a world-wide leader in the production of superbly engineered lubricants, additives and in-process and production cleaners.	DELCAM www.delcam.com Delcam is one of the world's leading suppliers of advanced CAD/CAM software product development solutions for the manufacturing industry.	DIAD Diad specializes in the storage, the treatment, and the recovery/recycling of dangerous and non-dangerous materials.
DORMER Dormer Tools is one of the world's largest producers of high speed steel cutting tools.	dti Working to create the conditions for business success and helping the UK respond to the challenges of globalisation.	FOOTPRINT SHEFFIELD Footprint Tools is one of the UK's leading privately owned manufacturers of hand tools serving the UK and world with tools using skills acquired from over 125 years experience.
GOODRICH The Goodrich Corporation company, is a global supplier of systems and services to the aerospace, defense and homeland security markets.	HAAS Haas is the world's largest CNC machine tool manufacturer. With a wide range of horizontal, vertical, lathes and turning centres. Haas has one of the highest service engineer to installation ratios of any UK machine tool distributors.	IFA IFA Ltd. are a specialist open die forging and ring manufacturer, producing forgings up to 10 tons in the widest possible range of metallic materials.
JOHNSON ALLEN Johnson Allen has been the market leaders in the supply of crack detection equipment for over 60 years.	LOLA Lola is a globally recognised supplier of specialist technology service and advanced composites to the motor-sport, automotive, aerospace, marine and communication sector.	MachineWorks MachineWorks Ltd. is the leading developer of NC simulation and verification technology. Their toolkit is the preferred solution for the leading software and hardware manufacturers in the world.
MAG CINCINNATI One of the world's leading suppliers of innovative, highly-productive machine tools.	MAHER Maher Limited is an international supplier of super-alloys, nickel and cobalt alloys, titanium, marriaging and special stainless steels and copper alloys in bars,	Materialise Industrial Services Materialise Group is best known for its activities in the field of rapid industrial and medical prototyping. It is also a leader in 3D printing,

steels,and copper alloys in bars, plate and forgings.		also a leader in 3D printing, digital CAD software and rapid manufacturing.
Messier-Dowty SAFRAN Group Messier-Dowty, a SAFRAN Group company, is the world-leader in design, development, manufacture and support of landing-gear systems.	METRIS Metris designs, develops and markets a unique range of 3D hardware and software inspection systems for the automotive and aerospace sectors.	Mettis Aerospace Mettis Aerospace is a leading provider of aerospace components based up world leading forging, machining and metallurgical skills Mettis Aerospace has a comprehensive range of manufacturing services on site.
MITSUBISHI ELECTRIC Mitsubishi Electric is recognized as a world-leader in the manufacture of electrical and electronic equipment used in information processing and communication.	Mitutoyo Mitutoyo was founded in 1934 aiming to be the first corporation to manufacture micrometers in Japan.	MLI High Performance Machining Specialists Manufacturing Laboratories Inc. specialises in high-performance machining and machine tools.
MORI SEIKI THE MACHINE TOOL COMPANY Mori Seiki produces extremely reliable machine tools and distributes worldwide.	NIKKEN Nikken Kosakusho produces some 500,000 toolholders, 3500 rotary tables and 250,000 reamers and maintains a reputation for producing quality products.	OPTIMA Optima Design Services is an authorised re-seller, certified training provider and leading provider of techincal support, and consultancy for PTC.
PTC The Product Development Company PTC specializes in product development to provide the best product lifecycle management and enterprise content management solutions.	Reliance Precision Limited Reliance produces precision motion control components and assemblies to exacting specifications in terms of both accuracy and performance.	RENISHAW apply innovation™ Renishaw's products provide manufacturers with the ability to machine components accurately, and perform measurement traceable to international standards.
Rolls-Royce Rolls-Royce is the world-leading provider of power systems in civil aerospace, defence aerospace, marine and energy.	Rotary Engineering UK Ltd is a world leading specialist in: Wind turbines, industrial electro-magnets, electrical workshop equipment, bespoke engineering solutions.	SANDVIK Coromant Sandvik Coromant is the world's-leading manufacturer of cutting tools for the metal-working industry, with more than 25,000 products.
starraghheckert the milling experts	TECHNICUT profit through performance	THIRD WAVE SYSTEMS Modeling technology for Machining Solutions

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Requirements and challenges

For Semantic KM

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Requirements: outline

- Effortless management
  - Penetration, affordability, reliability, reusability, agnosticism
  - Personalisation, accessibility use of paradigms familiar to users
- People, organisations, knowledge, services
- Ontological Coverage
- Support for Absolute Vs Local Knolwedge
- Autonomy and Coordination
- Identity, privacy and trust

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Challenges

- Reducing the amount of unstructured material
  - Empowering communities to formalise knowledge
  - To prepare for the future
- Extracting information from legacy data
  - To recover the past

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Challenges (ctd)

- Large scale of resources
  - Dozens of archives
  - Hundreds of thousands of documents
  - Several communities
- Cross-Media
  - Documents are cross-media
    - e.g. Powerpoint presentations, Word documents, HTML files
    - single medium does not contain all information
- Information integration over large scale
  - Fusion methodologies and techniques able to cope with uncertainty in fusion
- Focusing in a very large search space

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Challenges (ctd)

- Not all information and knowledge is to share
  - Personal knowledge
  - Departmental knowledge/Organisational knowledge
  - Public knowledge
- Capturing context of information
- Infrastructure
  - to support large scale (e.g. enabling sharing large resources)
  - support to discover and use of services

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Knowledge Acquisition

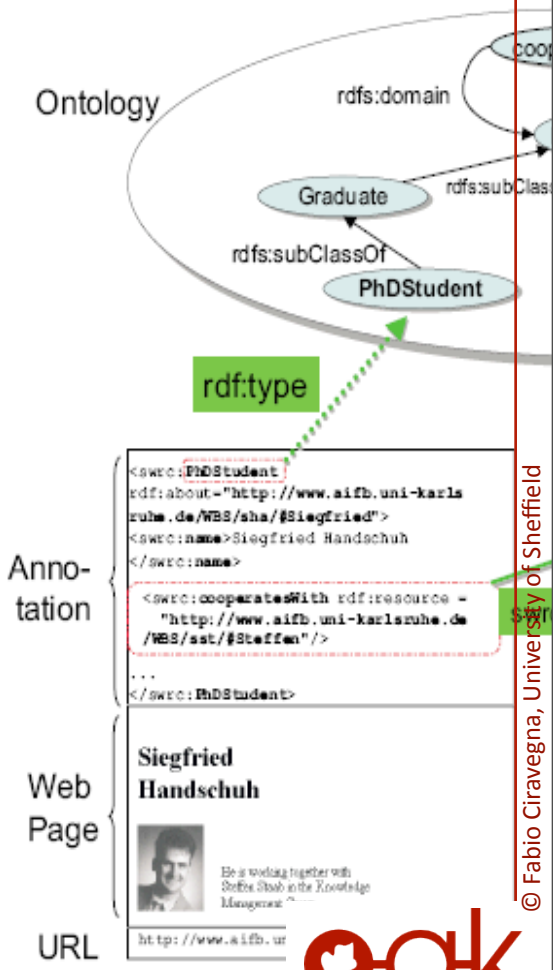
For Semantic KM

© Fabio Ciravegna, University of Sheffield

Ontologies and KM

Motivations for use:

- To represent the company's general view on the domain
  - How does the company work?
  - What is the company's official dictionary?
- As a middle layer to connect information from different information sources
  - The Web of data (as opposed to Web of documents)
- To represent communities' views of domains
  - e.g. marketing dept, customers, design and service departments have different views of the same products.
- Ontology mapping to navigate information sources
  - Mapping enables seamless communication among different worlds



© Fabio Ciravegna, University of Sheffield

The Semantic Web

- Instead of a few large, complex, consistent ontologies that great numbers of users share,
  - I see a great number of small ontological components
    - consisting largely of pointers to each other.
- Web users will develop these components in much the same way that Web content is created.

Jim Hendler: Agents and the Semantic Web, IEEE Intelligent Systems,No 2, Vol 16, 2001

© Fabio Ciravegna, University of Sheffield

Sem Web for Distributed Communities

- Enable freedom to communities
  - definition of community-specific views of the world;
    - capture and acquisition of knowledge according to them;

great number of small ontological components

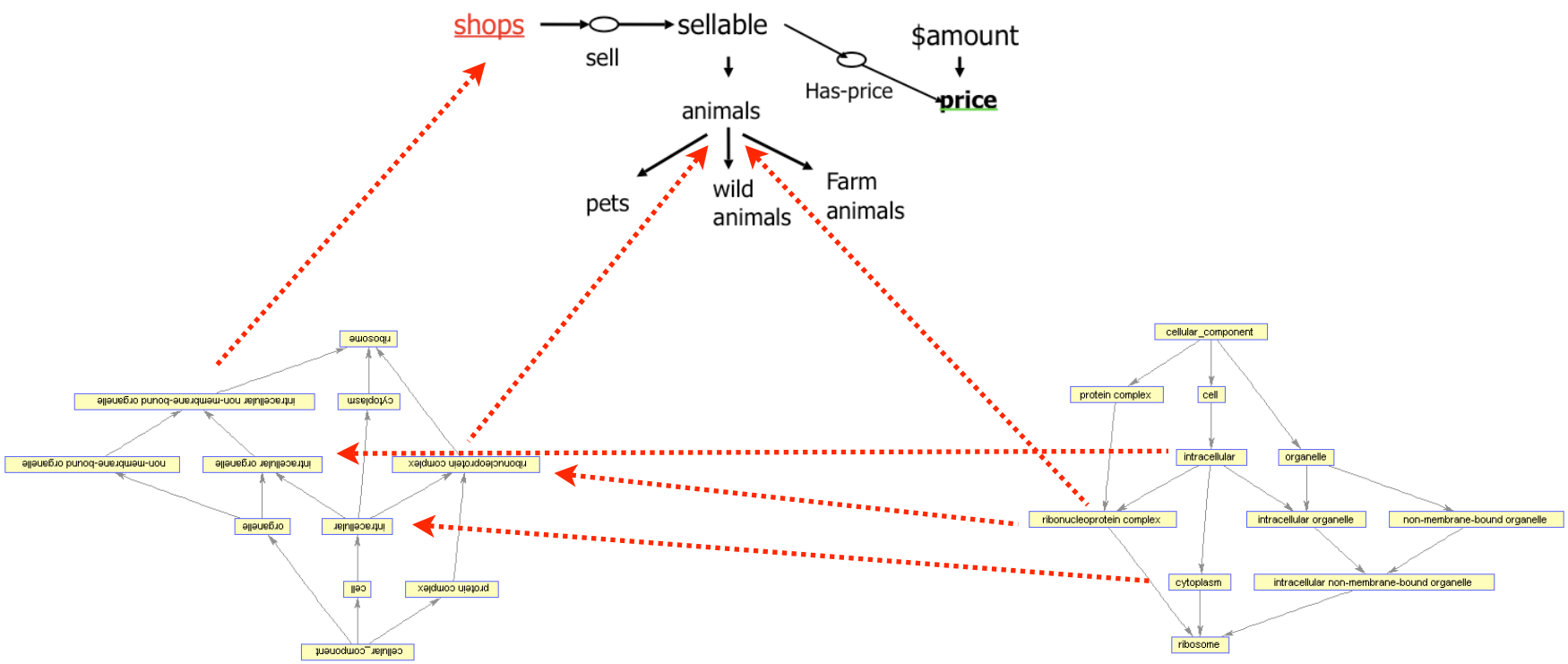
- Enable sharing with other communities
  - integration with the rest of the organisation's knowledge;
    - via integration of ontologies

consisting largely of pointers to each other

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

A web of ontologies



Knowledge can be

- shared using one ontology
- searched using another ontologies

© Fabio Ciravegna, University of Sheffield

Organisations, Information and Knowledge

The University Of Sheffield.

Requirements

- Reaching a critical mass of interconnections
  - comprehensive **enough** to enable operation
  - constantly growing
  - manageable in the long term
  - Not requiring central expensive control
    - Definition/maintenance
- It does not matter if connections are imperfect
  - a little semantics (integration) goes a long way

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

K-Form

Acquiring and capturing knowledge in a distributed environment

© Fabio Ciravegna, University of Sheffield

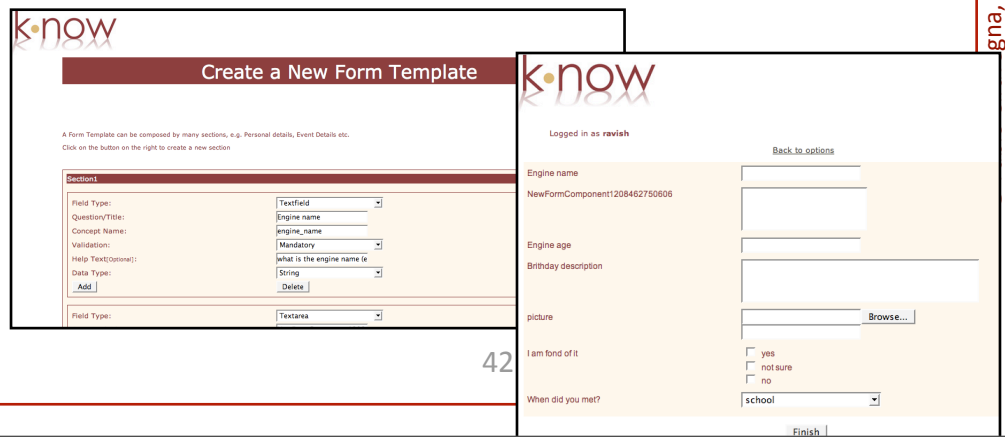
Ravish Bhagdev, Ajay Chakravarthy, Sam Chapman, Fabio Ciravegna, Vita Lanfranchi: Creating and Using Organisational Semantic Webs in Large Networked Organisations. 7th International Semantic Web Conference, Karlsruhe, Germany, October 2008.

The University Of Sheffield.

Development of ontological view

k-forms

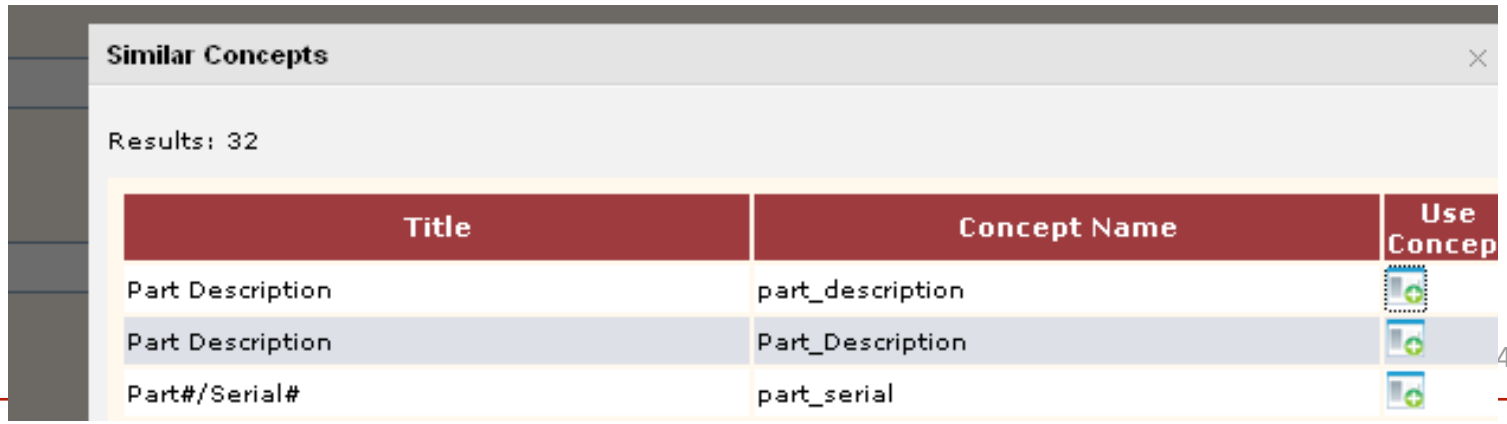
- Users define Web based forms visually using a Web browser
  - Forms, fields, type of information, possible values, etc.
  - Primitives are
    - Objects (e.g. engines), properties (e.g. no of cycles)
    - Objects can contain sub-objects as restrictions of a field
- Same freedom as Word/Excel forms
-



gna, University of Sheffield

Sharing among forms

- When form is created an ontology matching tool suggests relevant parts of other forms
  - to help users
    - create forms consistently without forgetting anything
    - reduce time
    - reduce time (saves specifying all details)
    - encourage sharing and linkage



© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Forms: Knowledge Capture

- When form is released users receive forms to fill
  - capture in local (no intranet connection)
  - upload to central repository in a second time
- Final Word/Excel document automatically generated
  - Can be read and printed and sent by email
    - as before
- Knowledge immediately available for search on the intranet

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

So what is the difference with Excel/Word/Google forms?

The Semantic Web

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Forms as ontologies

- The form schema is turned automatically into an explicit ontology
  - Objects are OWL concepts
  - Properties are
    - OWL properties if filler is base type
    - OWL Relations if filler is a nested object

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Linking ontologies

- When part of form is reused,
  - underlying ontology imports OWL concepts, relations and properties
- This **creates a semantic web of ontologies**
  - all the SW technologies used for managing distributed ontologies apply
    - e.g. distributed searching

- Users are told that importing is a serious matter
  - copy them otherwise

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Best Practice

- People tend to develop new forms starting from an existing forms
  - Hence the critical mass of connected ontologies is created

© Fabio Ciravegna, University of Sheffield

Organisations, Information and Knowledge

The University Of Sheffield.

Filling forms

- Semantic is assigned to the field values
- Filling forms creates RDF triples
  - Ontology-based search available
  - Different types of documents can be generated from the triples
    - to enable user tailoring
      - to adopt a specific terminology
      - for confidentiality

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Knowledge Capture

For Semantic KM

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Coping with Legacy Data

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Legacy Data?

- Legacy data in unstructured sources must be recovered
  - Access to knowledge captured with K-Forms must be seamlessly integrated with that extracted from legacy data (when possible)
- Requirement: extracting information from existing forms
- Method:
  - Use of automatic semantic annotation techniques
    - Natural Language Processing
    - Cross-Media Analysis

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

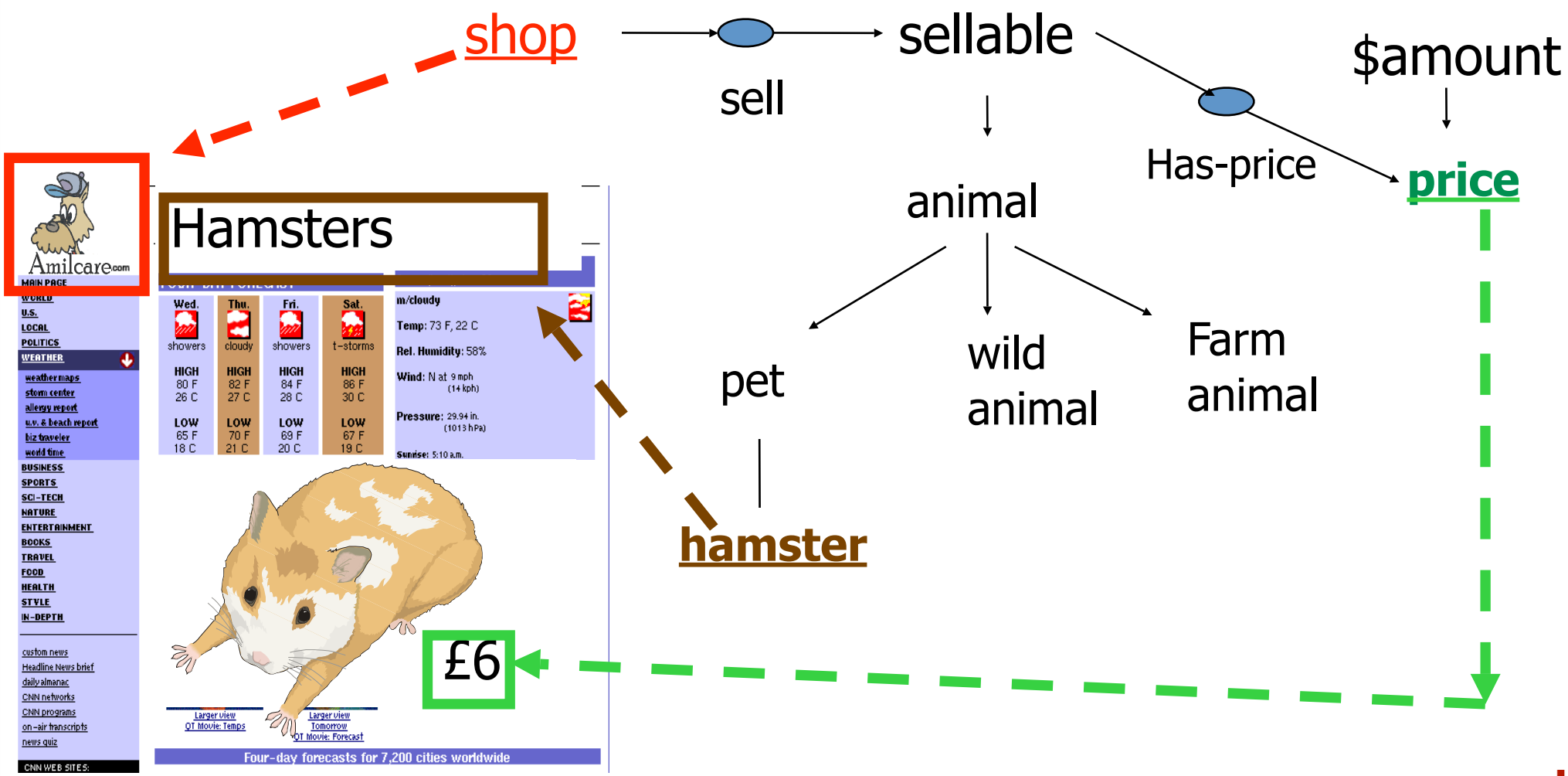
Ontology-based Annotations

- Marking up contained information
  - Portions of documents associated to objects in ontology
    - Allows:
      - Ontology-driven processing
      - Services based on ontology will be able to use information
    - Ontomat/CREAM (Staab et al 2001)
    - Melita (Ciravegna *et al.* 2002)
    - SemTag and Seeker (Dill et al. 2003)
    - ...and many others...

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Ontology-based Annotation



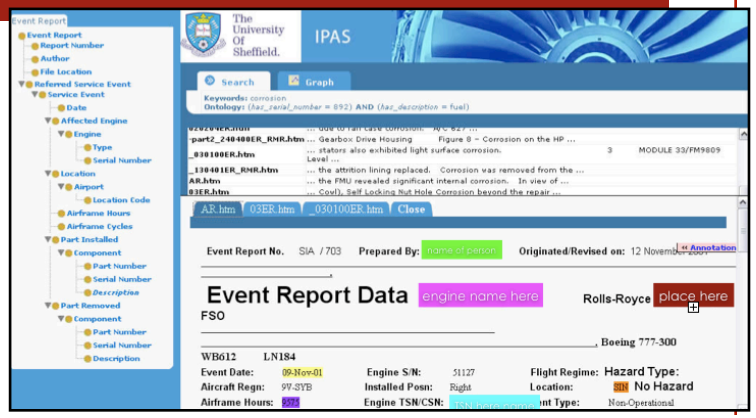
© Fabio Ciravegna, University of Sheffield

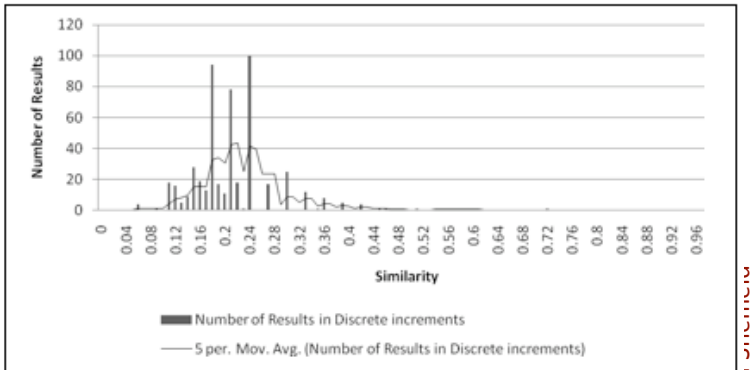
Organisations, Information and Knowledge

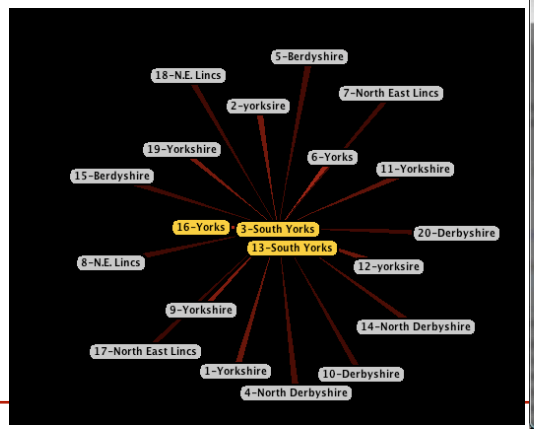
The University Of Sheffield.

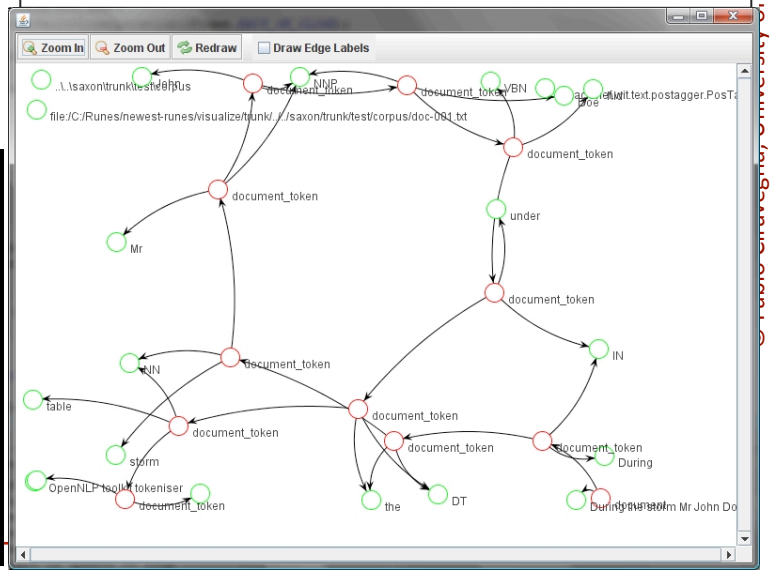
Capture from Legacy Data

- Tasks:
  - Text Categorization
  - IE
    - Entity Extraction [Greenwood 08]
      - Terminology Recognition [Butters 08]
    - Field Extraction [Ciravegna 08]
    - Relation Extraction
  - Information Integration [Chapman 08]









The University Of Sheffield.

Requirements

- Large Scale
  - In Number of archives
  - In size of archives
- Users cannot annotate documents
  - Annotation as side-effect of user actions
    - Task based (e.g. Filling forms with document content -> annotation)
  - Semi-supervised methods
    - Graph-based approach

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Annotation Engines

- Solutions like k-forms are very good for annotating new knowledge
  - large repositories of legacy data exist
  - it is important that new management solutions are able to reuse existing data
    - do not require a completely new world to be built for you!!
- Legacy data is generally represented in
  - databases
  - textual documents
  - images
  - ...

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Tasks for KA: Extraction

- Text:
  - Entity Extraction
  - Table Fields Extraction
  - Relation Extraction
  - Event Extraction
- Data:
  - Similarity of Data Instances
  - Functions and relation
  - Finding patterns and (ir-)regularities in data
- Images:
  - Semantically driven Image analysis using ontologies, for retrieval and annotation
  - Image classification/ clustering with respect to the dominant visual trends

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Information Extraction

- Automatically extracting pre-specified information from textual documents
  - salient facts about pre-specified types of events, entities or relationships.
- Populating a structured information source from a semi-structured, unstructured, or free te

WASHINGTON, D.C. (October 5, 1999) -
nQuest Inc. today announced that Paul Jacobs, for
Vice-President of E-Commerce at SRA Internatio
has joined the company's executive management t
as president.

Company: nQuest Inc.
Date: today
InPerson: Paul Jacobs
InRole: president

Company: SRA International
OutPerson: Paul Jacobs
OutRole: Vice-President of E-Commerce,

© Fabio Cia

Organisations, Information and Knowledge

The University Of Sheffield.

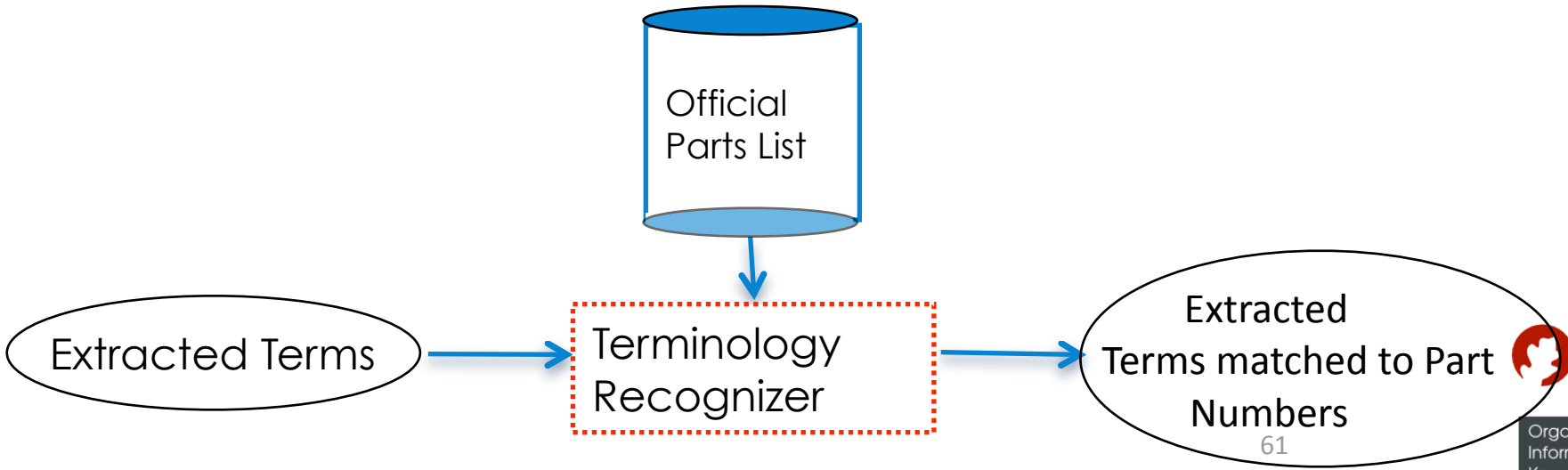
Named Entity Recognition

- Tasks:
  - Recognition and classification of named entities
    - E.g. people's names, companies, locations, etc.
  - Unique identification of named entities (URI assignment)
    - Including disambiguation
      - Michael Jordan as basketball player Vs lawyer
      - London UK Vs London USA
  - Integration with other sources
    - E.g. positioning on a map

© Fabio Ciravegna, University of Sheffield

Terminology Recognition

- NER is one example of term recognition
- More useful in technical domains is terminology recognition
  - The task of assigning a URI to a technical description
    - i.e. mapping a natural language description to the official company ontology



© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

More complex IE: event modelling

- Not just NER but also relation among elements in a document
  - More complex task
  - Requires some reasoning to bridge the complexity of events to the ontology structure
    - Imprecision in extraction
    - Information non matching the ontology schema
- This is where IE has hit a performance ceiling
  - 60/70 Precision/Recall ratio since 1998

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Table Field Extraction

- Tables are an essential part of many documents
  - Most information is represented in tables
- Tables can be represented as forms to fill
  - Semantics is fixed
  - Wrapper writing or wrapper induction (Kushmerick 1997)
- Tables can be created ad hoc in documents (e.g. Word docs)
  - Semantics is unclear
  - Sometimes documents are created as part of a workflow, therefore they tend to be created using common models
    - e.g. by re-using the previously generated document
    - hence tables evolve, but still semantics can be traced

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Information Integration

- Facts from different sources need to be integrated
  - To connect information/knowledge across docs
    - Assign unique URI
  - To solve discrepancies and ambiguities
- Steps
  - Unique instance identification (for entities)
  - Record linkage (for events)
- Information Integration strategies
  - Generic
    - Distance metrics (Chapman 2004)
    - Using Web bias
    - Statistical matching
    - Application specific
      - Rules

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

SimMetrics

- Library of distance metrics released as open source
  - http://sourceforge.net/projects/simmetrics/
  - >20,000 downloads since end of 2004
  - Most downloaded distance metrics library on the Web
    - for strings and records
  - Hundreds of applications
  - Developed by Sam Chapman, University of She

Anastasiosyal.com

So what's the buzzword today!?

<< SQL Server script to auto-create indexes on all Foreign Key Columns

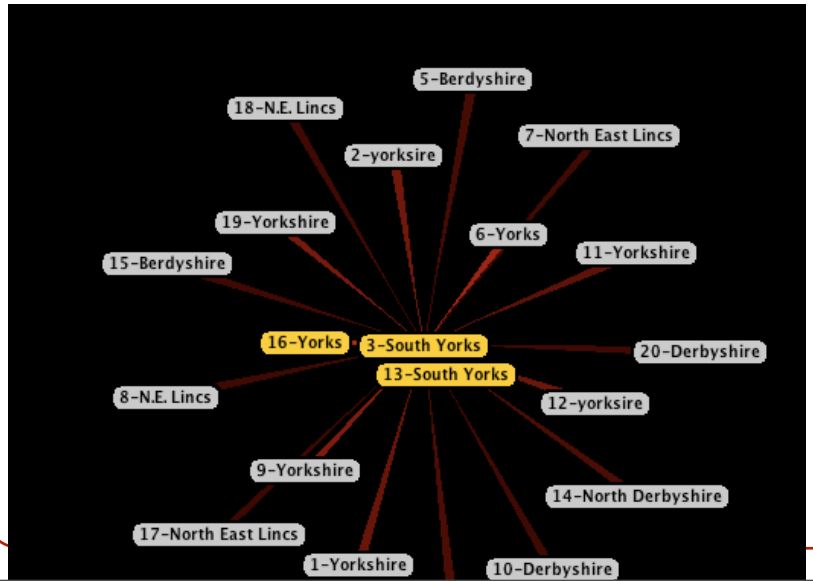
Beyond SoundEx - Functions for Fuzzy Searching in MS SQL Server

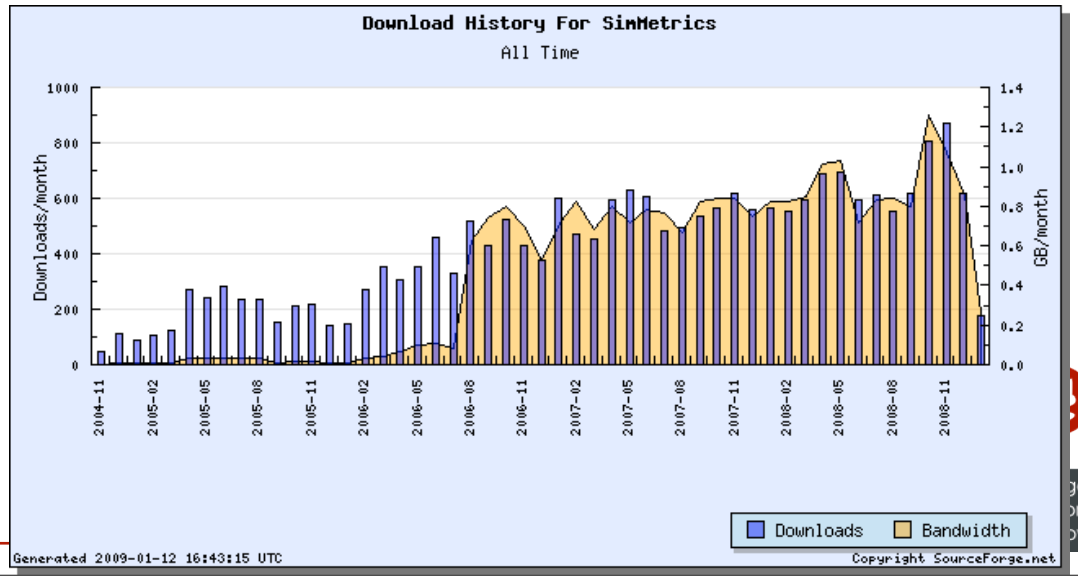
Sunday, January 11, 2009 12:08 AM, Filed Under SQL C#

In this post:

1. SoundEx in Sql Server
2. SimMetrics
3. Adding string Metric functions in MS Sql Server
4. Evaluating metric accuracy and comparing Metrics
5. Conclusion + code

Quite often we come across a requirement where we may need to perform some sort of fuzzy string grouping or data c
customer records of a database by identifying records that are similar but not necessarily exactly the same (due to spell
not successfully group such data. We will need to employ what is commonly referred to as a distance algorithm or a stri
are.





© Fabio Ciravegna, University of Sheffield

ganisations, rmation and owledge

The University Of Sheffield.

Armadillo: Historical Data Mining

http://www.hrionline.ac.uk/armadillo/

The Marine Society Registers

The Westminster Historical Database

Eighteenth Century Fire Insurance Policies

Prerogative Court of Canterbury Wills

The Proceedings of the Old Bailey

AHDS Deposits

St. Martin's Settlement Exams Index
WESTCAT

Collage image databse
Guildhall Library

Harben's Dictionary of London

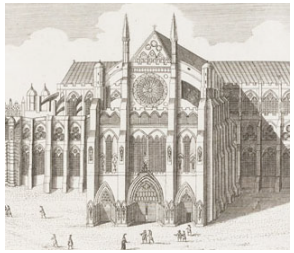
John Strype's "Survey…"

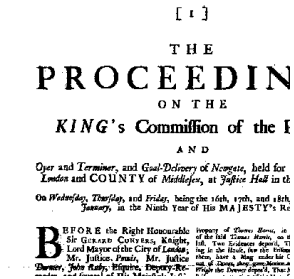
http://www.motco.com

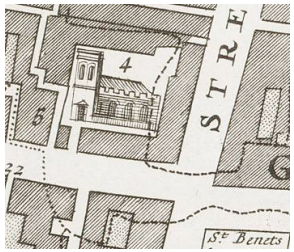
Metropolitan London in the 1690s
IHR

Selected Criminal Records
PRO

House of Lords Journals
BOPCRIS







Arts & Humanities Research Council

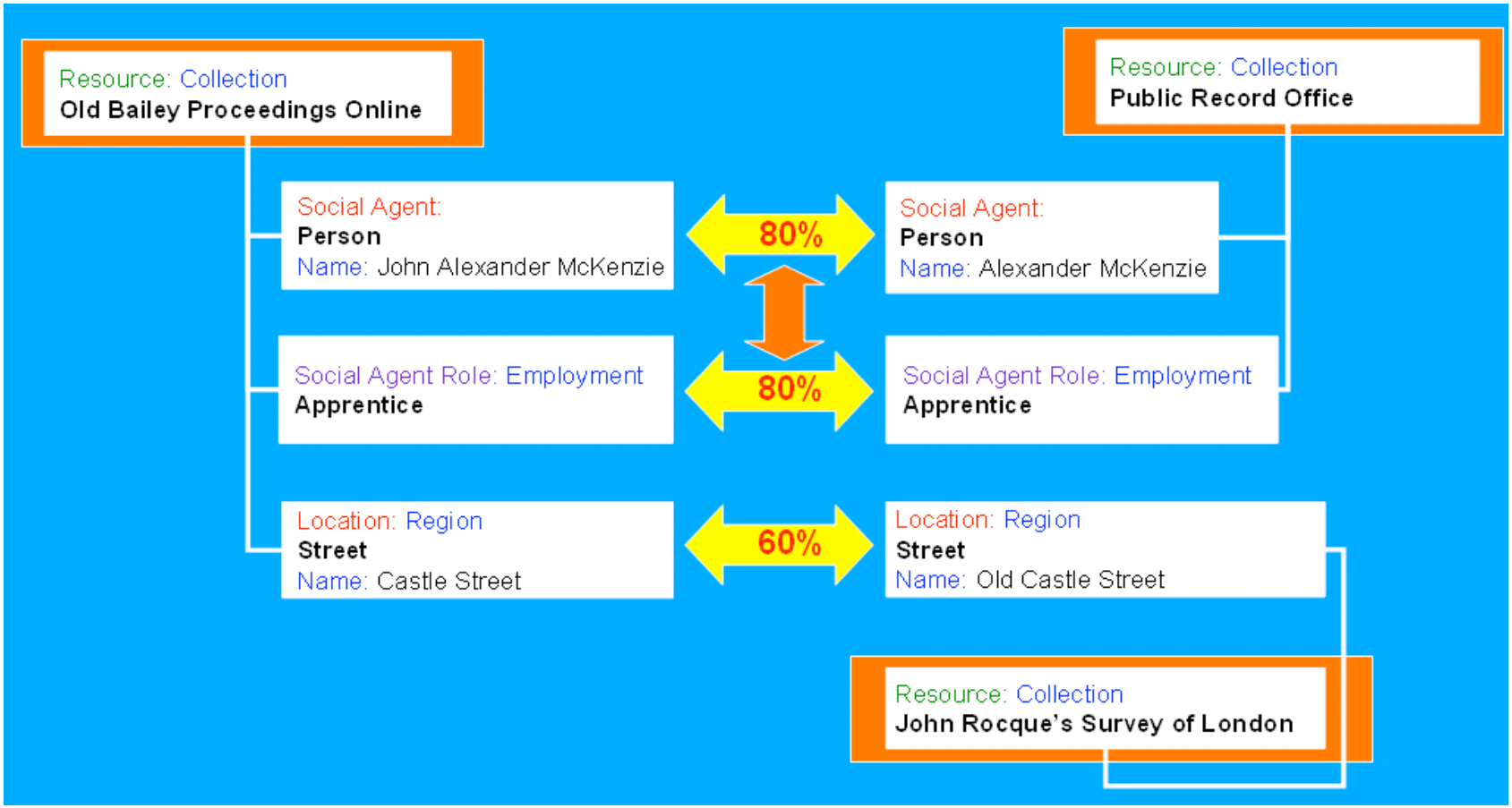
Organisations, Information and Knowledge

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Information Integration

Armadillo: Historical Data Mining



Arts & Humanities Research Council

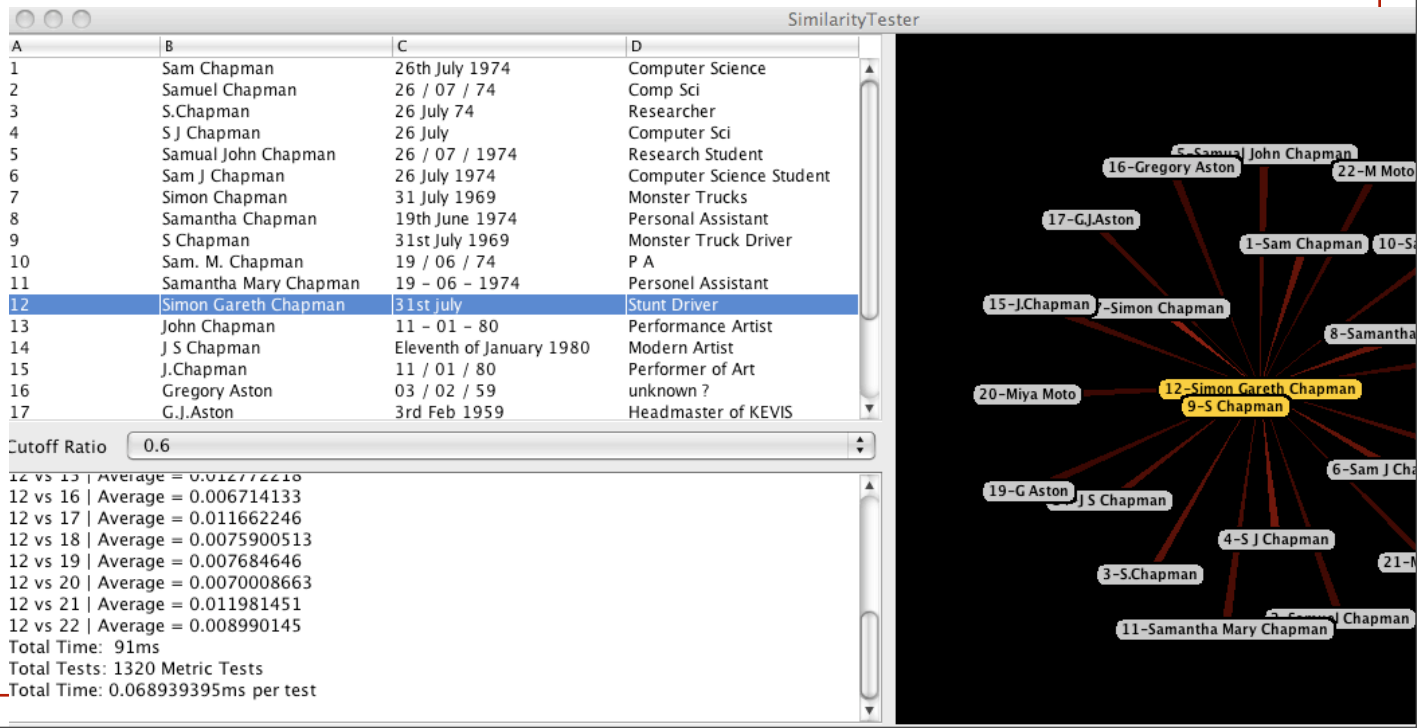
© Fabio Ciravegna, University of Sheffield

Organisations, Information and Knowledge

The University Of Sheffield.

Department of W&P Appl

- Goal: identifying the most likely record matching a NL description
  - I want to become (formation)
    - A plumber (job)
    - In South Yorkshire (geography)
  - Given misspellings and lack of facility with computers
    - “become plumba in Sothyorks”



The University Of Sheffield.

Sharing Knowledge

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Challenges (ctd)

- Not all information and knowledge is to share
  - Personal knowledge
  - Departmental knowledge/Organisational knowledge
  - Public knowledge
- Capturing context of information
- Infrastructure
  - to support large scale (e.g. enabling sharing large resources)
  - support to discover and use of services

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Knowledge Retrieval

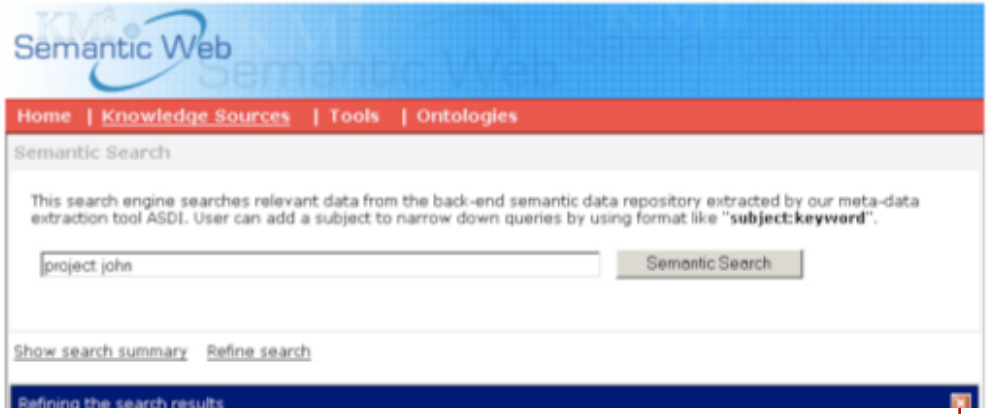
for Semantic KM

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Approaches to Semantic Search

- Keyword-based approaches
  - Require translating all the keywords in order to perform the query
    - E.g. SemSearch

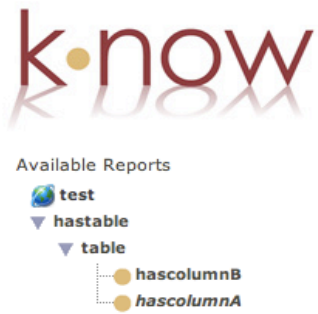


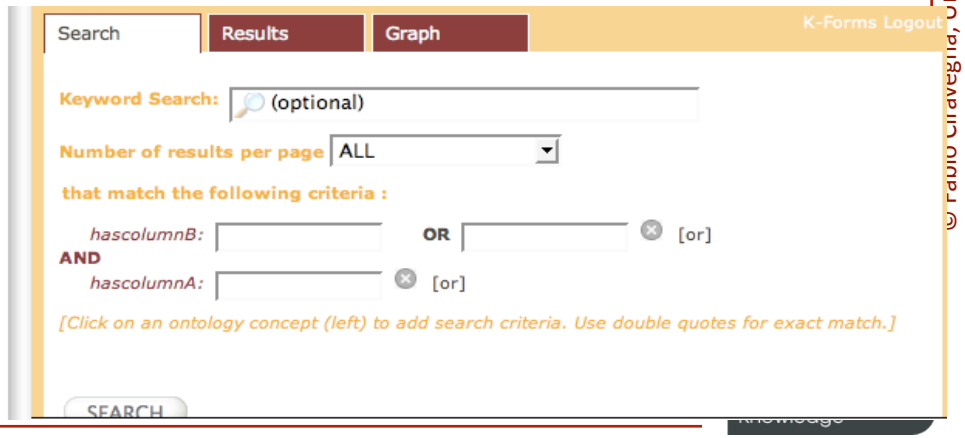
- View-based approaches
  - Based on querying by building visual graphs
    - E.g. Falcon

Search Strategy (ctd)

- A natural language approach
  - E.g. Aqua

- Form-based approacnes
  - e.g. k-search





© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Ontology-based Querying: Issues

- Metadata can cover just part of the material of interest to the users
  - The information not annotated using metadata is irretrievable
- How many topics can we model with Information Extraction?
  - 21 topics/ 14 topics partially or not covered by annotations
    - given size of corpus there is no way that manual annotations are added

- Often the use people will do of information is impossible to foresee
  - communities organise forms for themselves
  - some information not structured
    - text fields
- Sometimes Information is impossible to retrieve reliably using automatic methods
- If automatic means are used, often some parts of the knowledge is beyond the current technical capability

The University Of Sheffield.

Issues and Solutions

- Ontology can be extended
  - But increases effort in indexing
    - Equivalent to extending metadata in SDM
  - But it is impossible to foresee all uses of information
    - Ontology will always be insufficient somehow
- Information Extraction can be used to reduce burden of annotation
  - But some parts are irretrievable

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Hybrid Search (keywords+ontology)

- Mixes keyword and ontology based search
  - Ontology based search
  - Traditional keyword search
  - Keyword in contest of ontology-based annotations

affected parts is concept in ontology

- Potential queries:
  - Return all documents where the word fuel is mentioned
  - Return all documents where the affected part description includes the word fuel
  - Return all documents where the affected part description is similar to “fuel duct”
  - Return all documents where the affected part description is equal to “fuel duct” (URI=XXXXX)

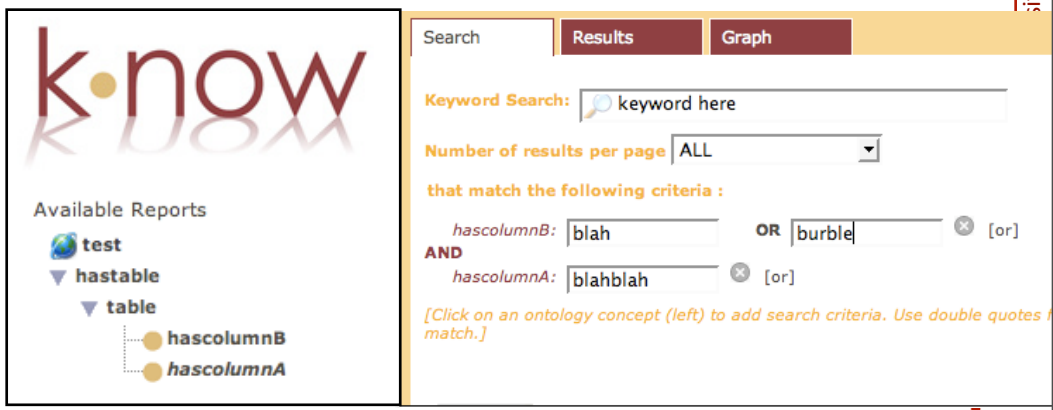
© Fabio Ciravegna, University of Sheffield

Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Vitaveska Lanfranchi and Daniela Petrelli:
Hybrid Search: Effectively Combining Keywords and Semantic Searches, in Proceedings of the 5th European Semantic Web Conference, ESWC 08, Tenerife, June 2008

The University Of Sheffield.

Hybrid Search

- Keywords and ontology-based search can be mixed within the same query
  - Pure ontology-based searching
    - When metadata covers information precisely
  - Keyword-in-context of annotation
    - To match strings in text annotated with semantics (textual form fields)
      - e.g. “fuel” is matched only on snippets of texts annotated as removed parts
  - General Keyword querying
    - For searching on the document/form as a whole

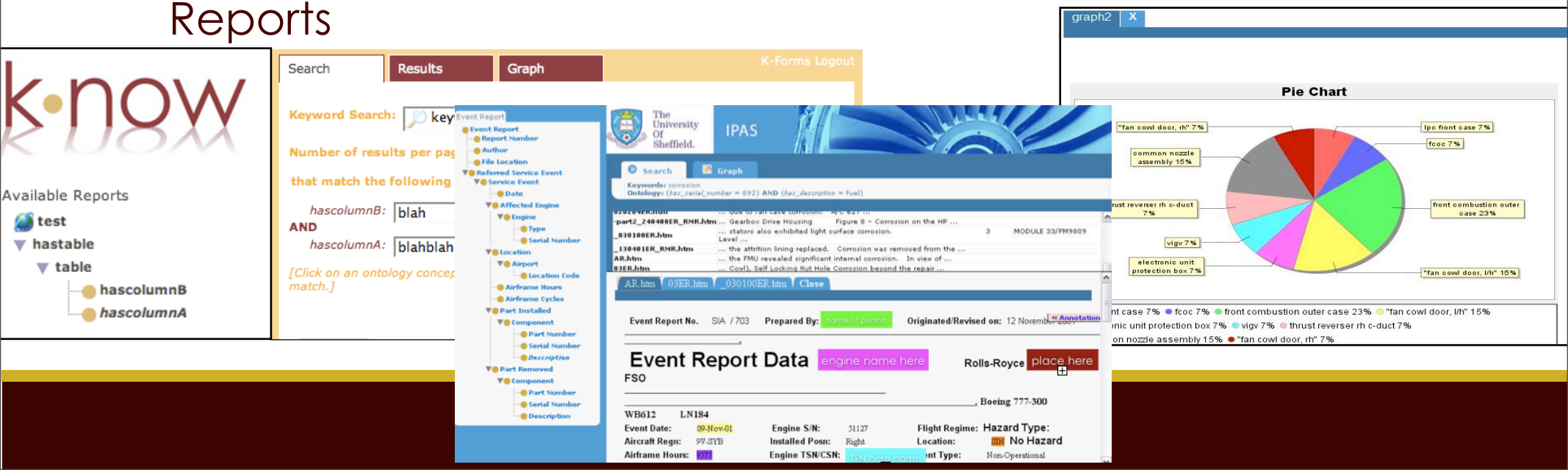


ity of Sheffield

Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Vitaveska Lanfranchi and Daniela Petrelli: Hybrid Search: Effectively Combining Keywords and Semantic Searches, in Proceedings of the 5th European Semantic Web Conference, ESWC 08, Tenerife, June 2008

K-Search

- Enables querying documents using hybrid search
- Enables quantification of unstructured information
- Currently applied at Rolls-Royce to:
  - Event Reports (1998-2004), Technical variances, Module condition reports, Strip reports, ERMS, Event Summary Reports



The University Of Sheffield.

Sharing and Reusing Knowledge

- Distributed interconnected resources
  - can be queried across via interconnected ontologies
    - Searching metadata rather than text
  - Retrieving information independently from the store/ media
  - Enables querying resources using my ontological view
    - largely independently from the view used originally to create it

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

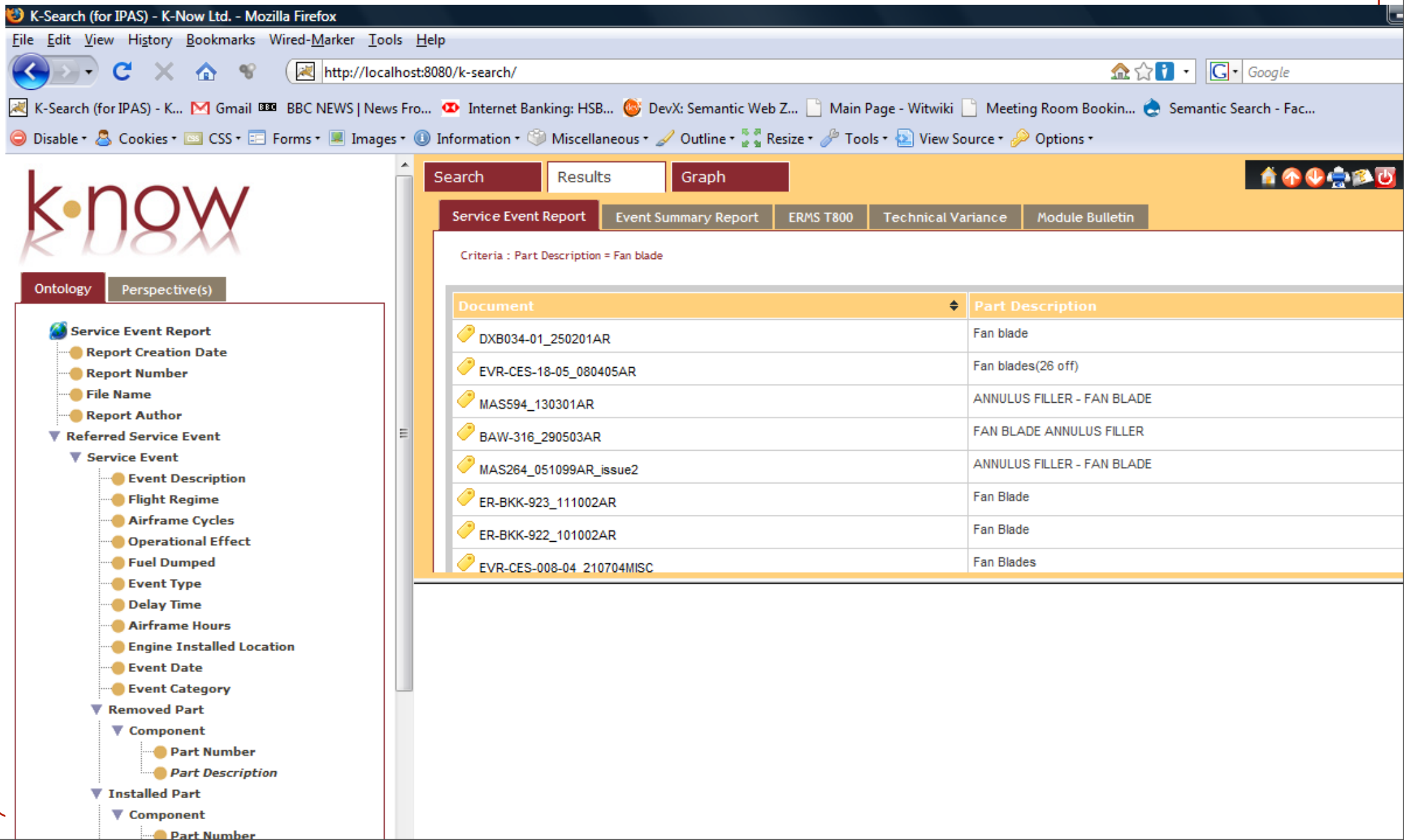
Support for dynamic communities

- K-Search enables searching multiple repositories at once using one of the available ontologies
  - Query a specific resource via the original ontology
  - Query a resource using a different ontology interconnected to the original one
  - Query multiple repositories using one specific ontology.

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Searching Across



The University Of Sheffield.

Support for dynamic communities

- When an ontology different from the original is used
  - the original query is mapped to the original ontology via the formal links.
  - For the parts that are not mapped the restrictions are turned into keywords

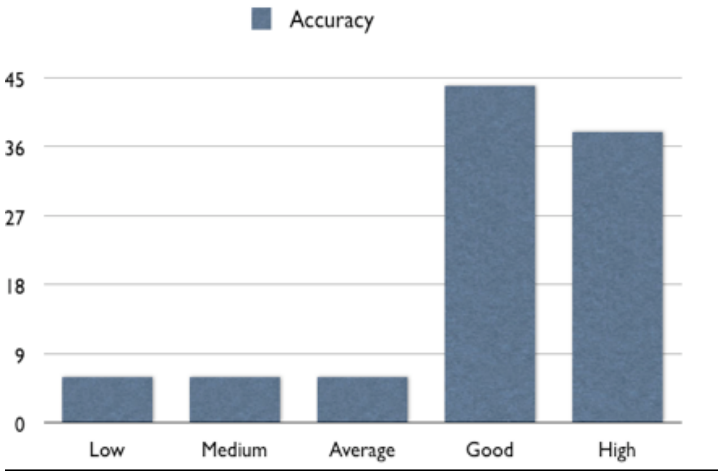
© Fabio Ciravegna, University of Sheffield

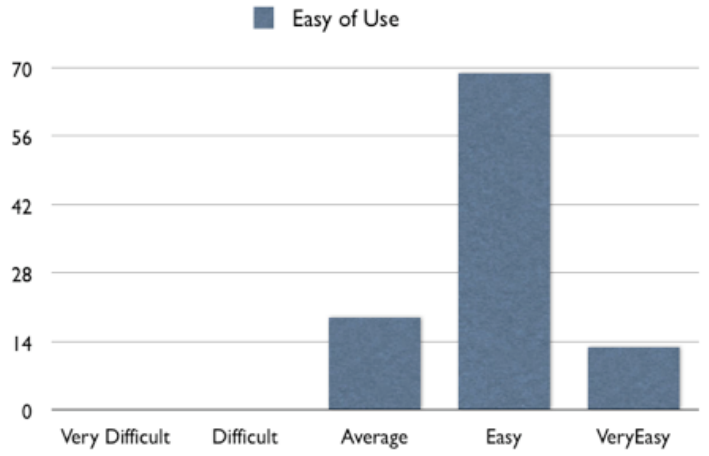
The University Of Sheffield.

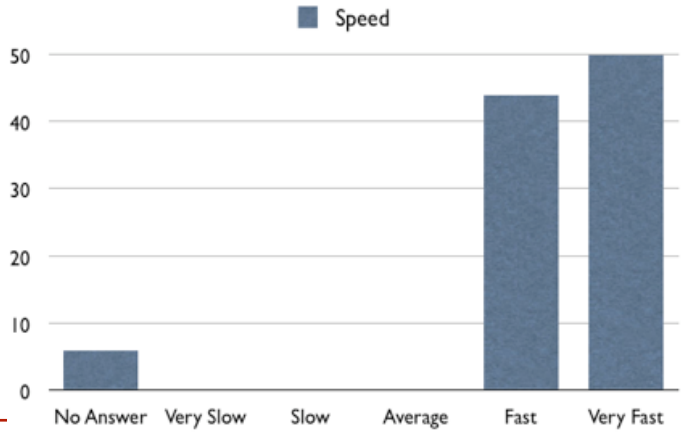
Liked by the users?

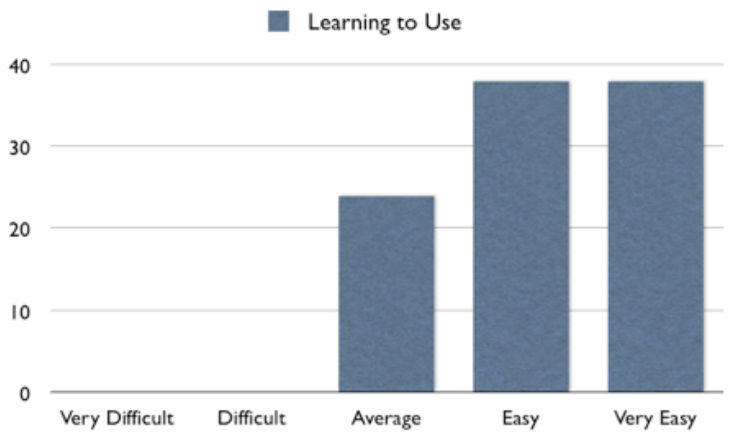
- K-Search + IE
  - Finalist of Rolls-Royce Creativity Award 2007
    - Voted by employes for its innovation potential

k-now.co.uk









© Fabio Ciravegna, University of Sheffield

Organisations, Information and Knowledge

Liked by Users?

- Developed as part of IPAS collaboration
  - 2005-2008
    - £240,000 (50% Rolls-Royce)
- Support to the design of new Trent XWB
  - Porting to 9 Information Sources
    - 2008-2009
    - around £100,000 (100% R-R)

- Funds from Rolls-Royce for use of K-Tools for use in manufacturing
  - around £340,000

© Fabio Ciravegna, University of Sheffield

Other Applications

- TalkBack: Customer Retention in Gyms

- DU360: adaptive information integration for products/vendors from different suppliers and EU standards
  - Several large customers in detail commerce (supermarkets)

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Knowledge Visualisation

In Semantic KM

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Understanding Knowledge

- Semantic technologies enable representation of content in a uniform way
  - Independently from the way data was originally expressed
    - e.g. documents, images, etc.
- Great for reasoning, not for people
- Users need another form
  - E.g. accessing the original source (i.e. the text or image it comes from)
  - Via some easily graspable visualisation means, i.e. facts graphed according to space and time

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Visualisation

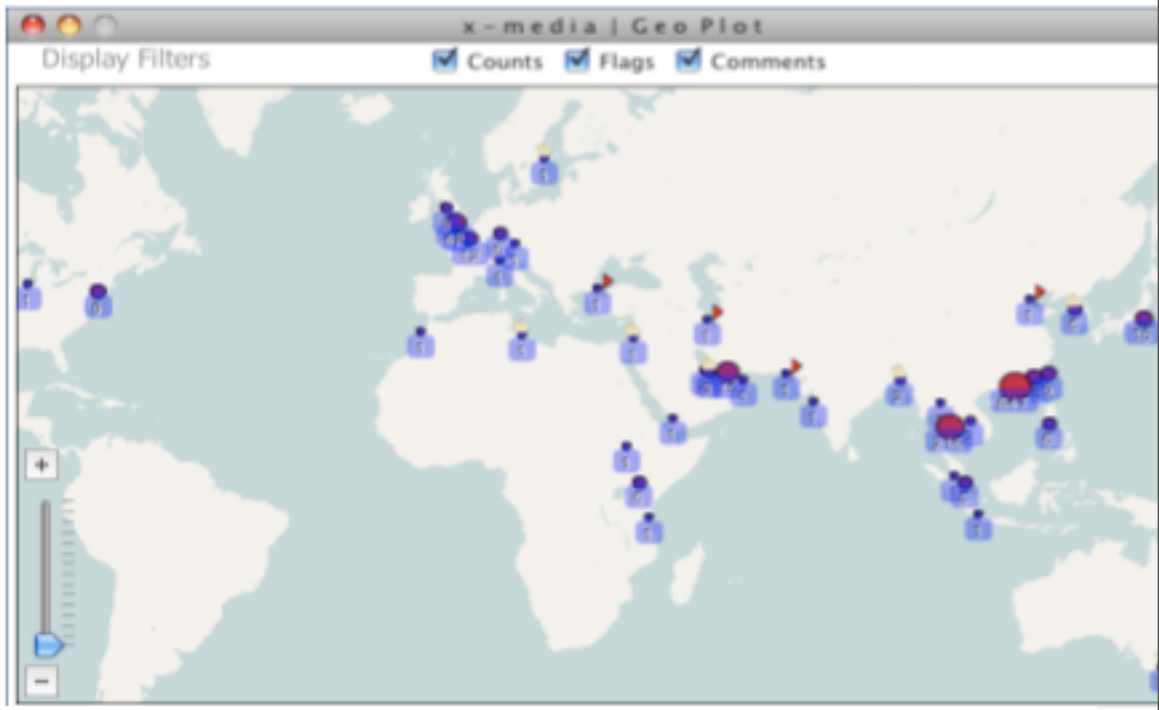
- “the Semantic Web emphasises formal, machine readable […] approaches. It focuses on the formal and even the meaning achieved through rigorously defined forms.
- Information visualization emphasizes the semantics and the meaning that can be conveyed by visual-spatial models to the users.” [6].

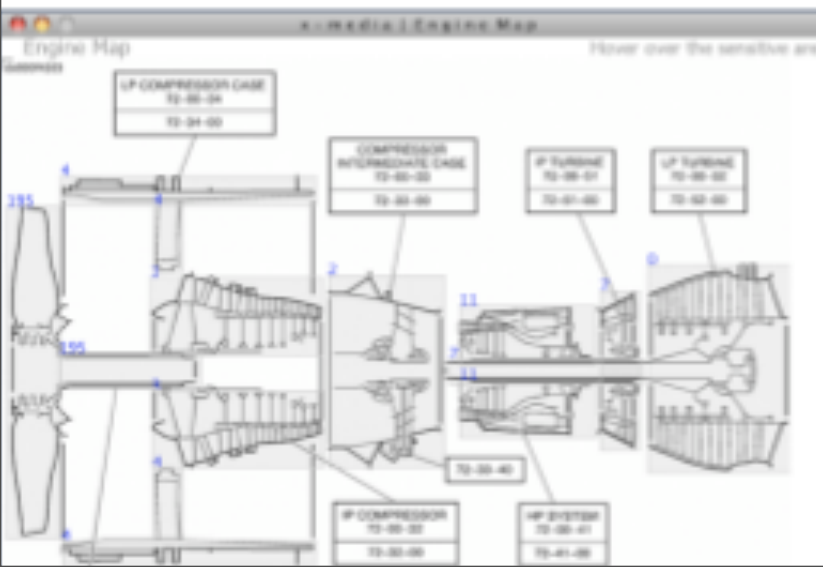
© Fabio Ciravegna, University of Sheffield

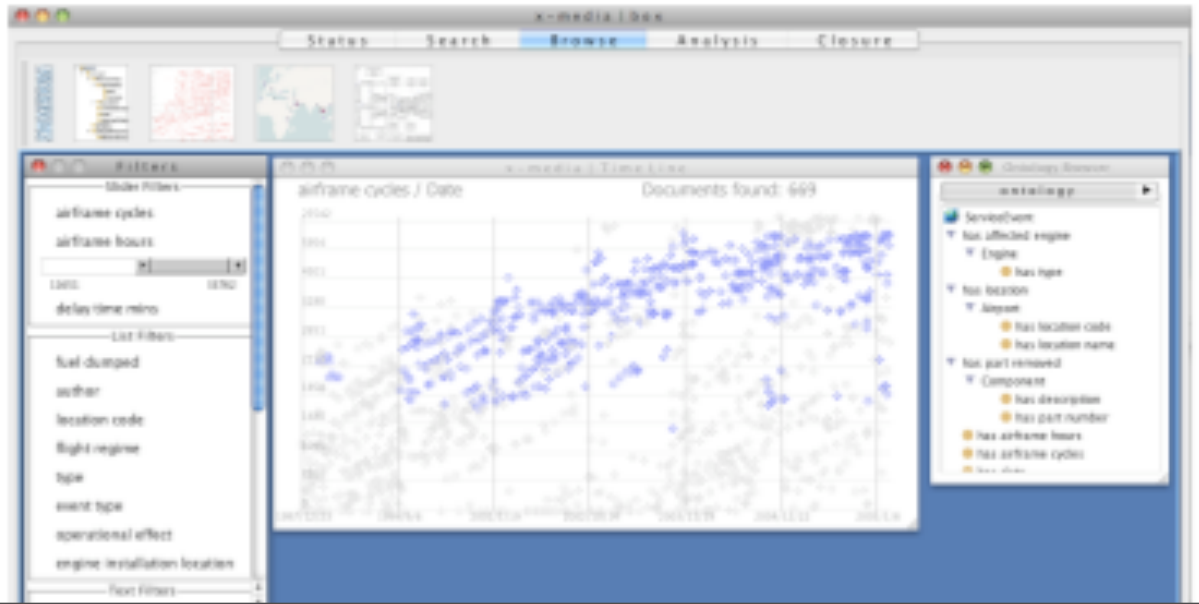
The University Of Sheffield.

3 level interconnected views

- Geographical
- Topological
- Timeline
- With dynamic filtering
  - Petrelli, Mazumdar, Dadzie, Ciravegna Multi Visualisation and Dynamic Query for Effective Exploration and Annotation of Semantic Data







© Fabio Ciravegna, University of Sh

The University Of Sheffield.

Reusing Knowledge

In Semantic KM

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Steps in Reuse

- Set up:
  - the sense-making activity is triggered by the need for new knowledge.
  - Evidence in different forms and from all sources available must be collected in order to clearly define the analysis and the exit conditions for the sense-making activity – when the user can actually "make sense" of the situation.
- Knowledge gathering:
  - semantic and similarity search support knowledge retrieval from the semantic repository, supported by ontology-based exploratory information retrieval.
  - The knowledge extracted: document content and metadata describing provenance and uncertainty, is used in subsequent phases.

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Steps (ctd)

- Knowledge analysis and manipulation:
  - support is required for interacting with knowledge, and for capturing the content and context of implicit information (e.g., comments, open requests and action lists) via ontology-based annotations, in order to create new, explicit knowledge.
- Recording of the current status:
  - sense-making may span an extended period. We propose levels of abstraction that record its status from a high level using summaries, to the level of knowledge analysis – the visualisation of islands of knowledge and their ontological relations, to the detail and context of the knowledge – information provenance and co-occurring facts.

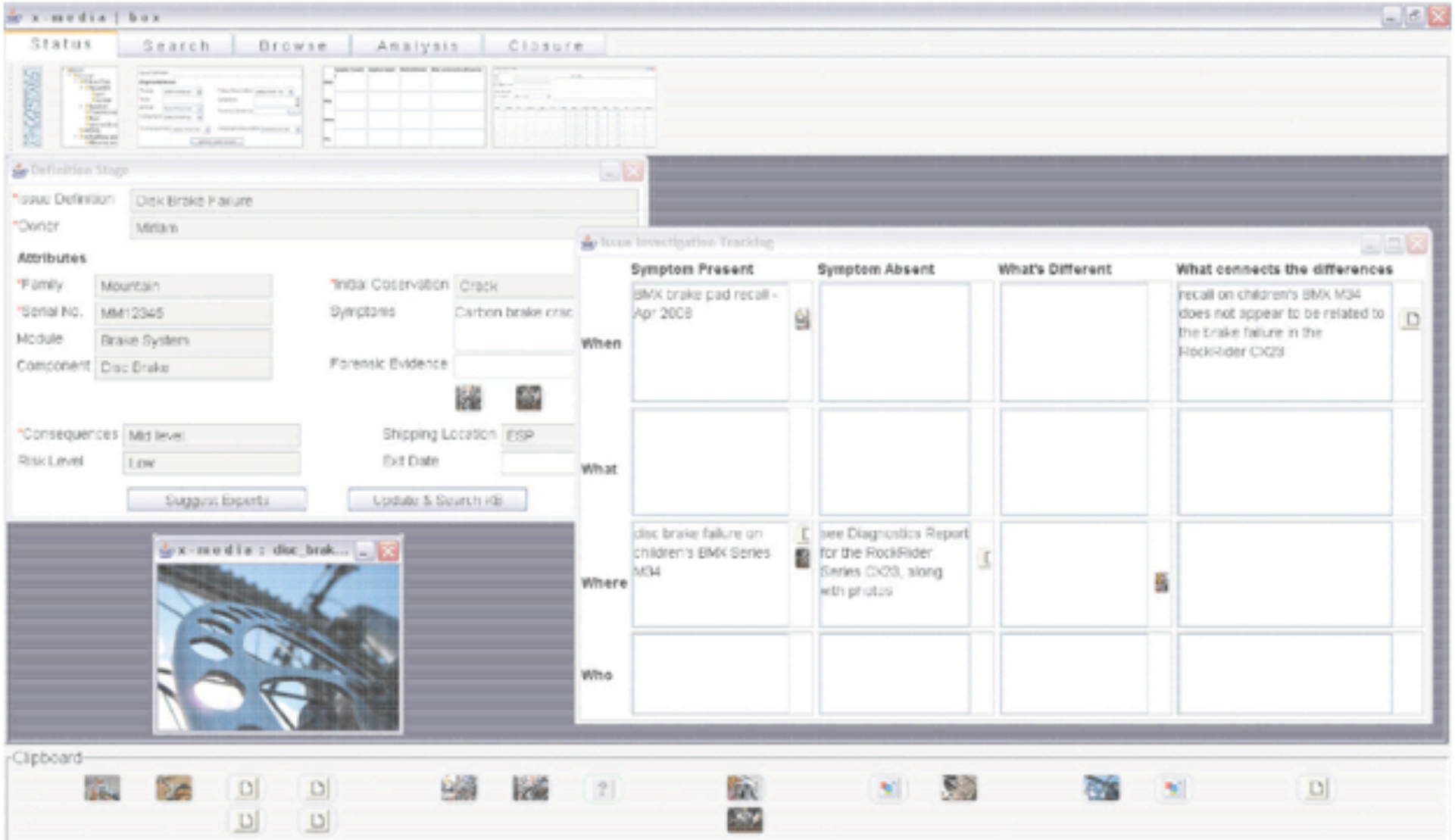
© Fabio Ciravegna, University of Sheffield

Example: XMediaBox

- Set up a new case
- The Search tab provides different search capabilities
- The Browse tab enables browsing of the user's individual workspace and the knowledge repository
- The Analysis tab contains the knowledge lenses
- Finally, the Closure tab is used to record
  - the extent to which the sense-making activity has satisfied the exit criteria,
  - the "lessons learnt", in order to support sharing and reuse.
- The Semantic Clipboard in the lower part is used to move data and knowledge elements between tools or temporarily store resources.

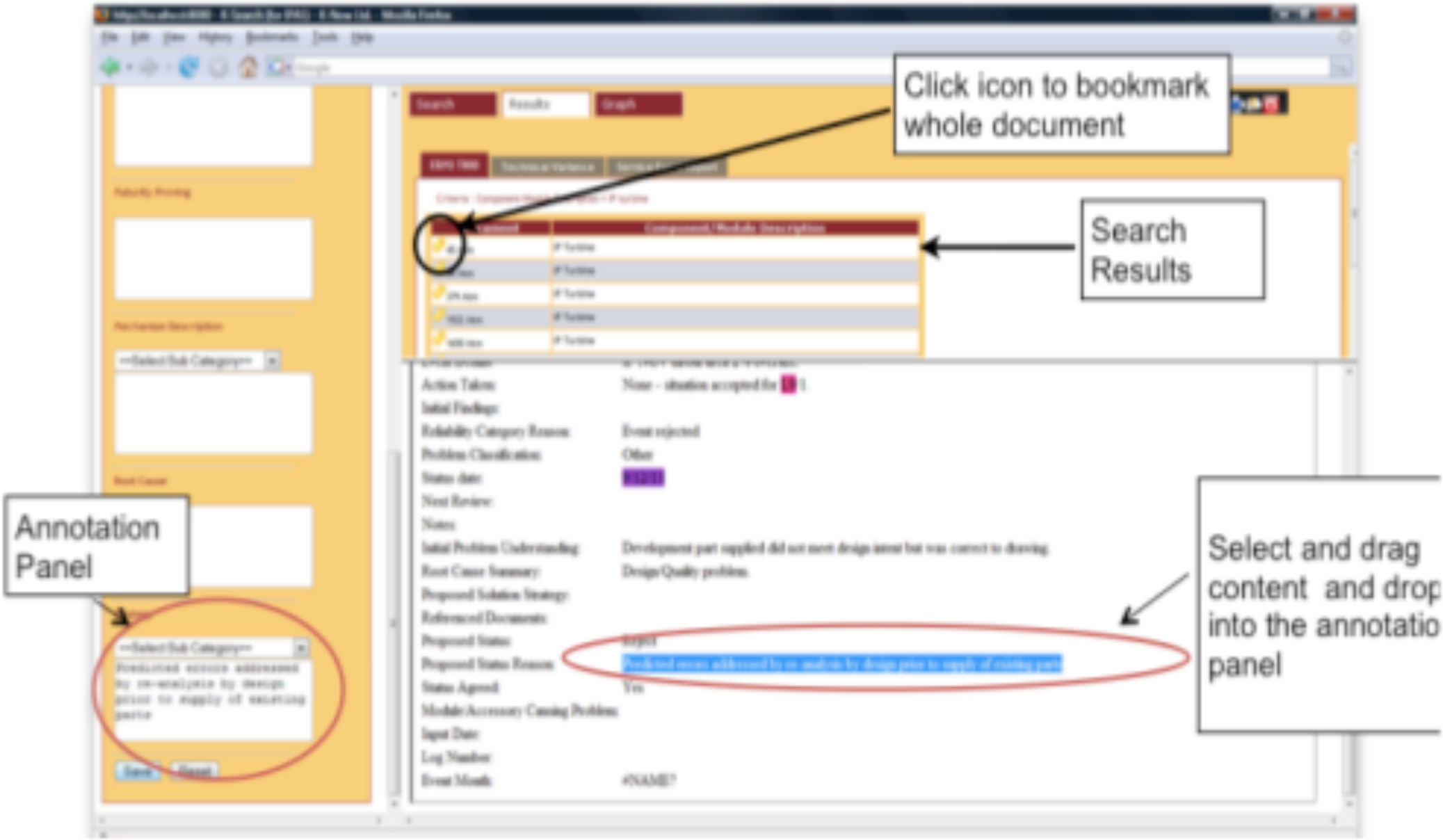
© Fabio Ciravegna, University of Sheffield

Set Up



© Fabio Ciravegna, University of Sheffield

Searching & Annotation



Search Tab

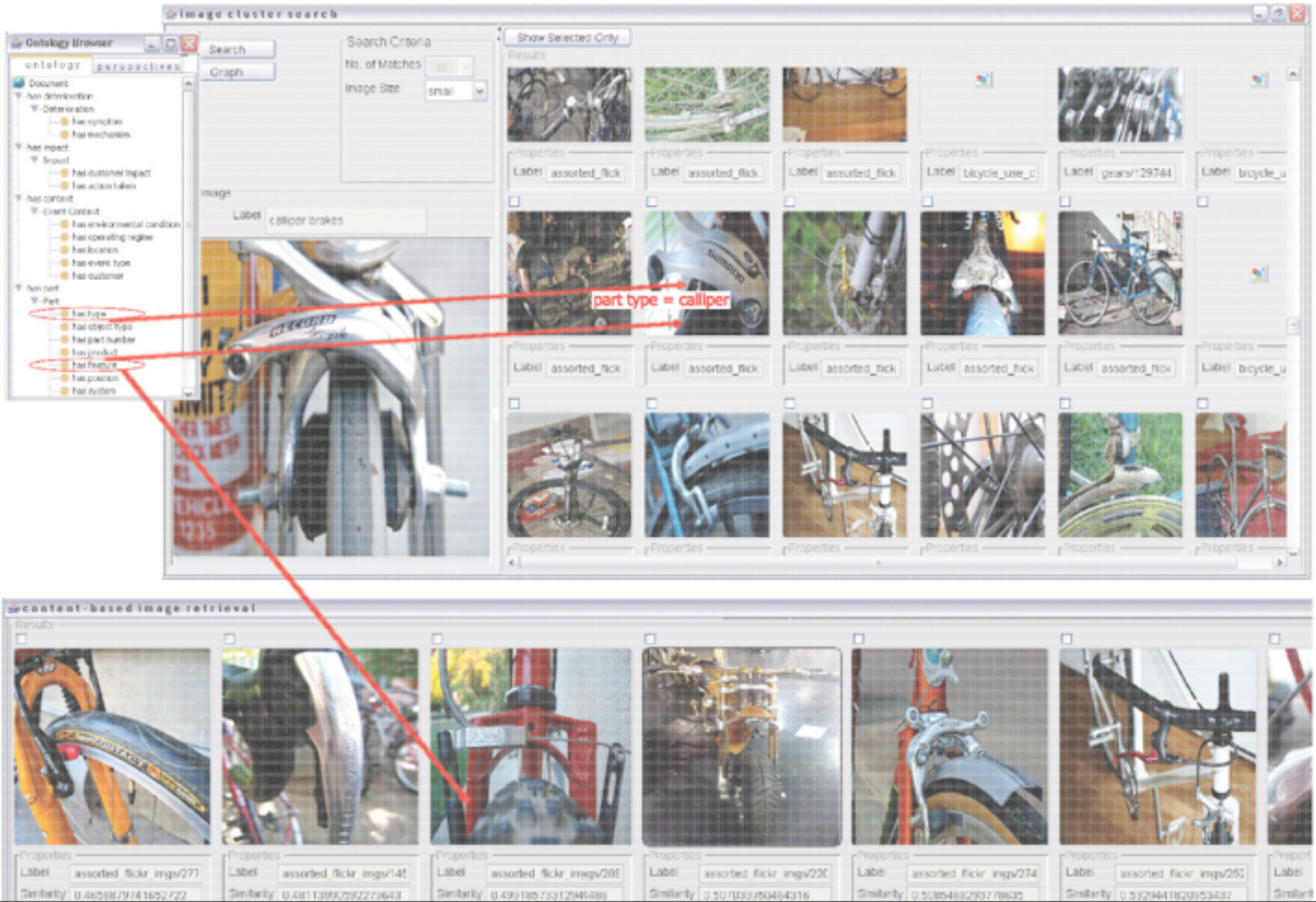
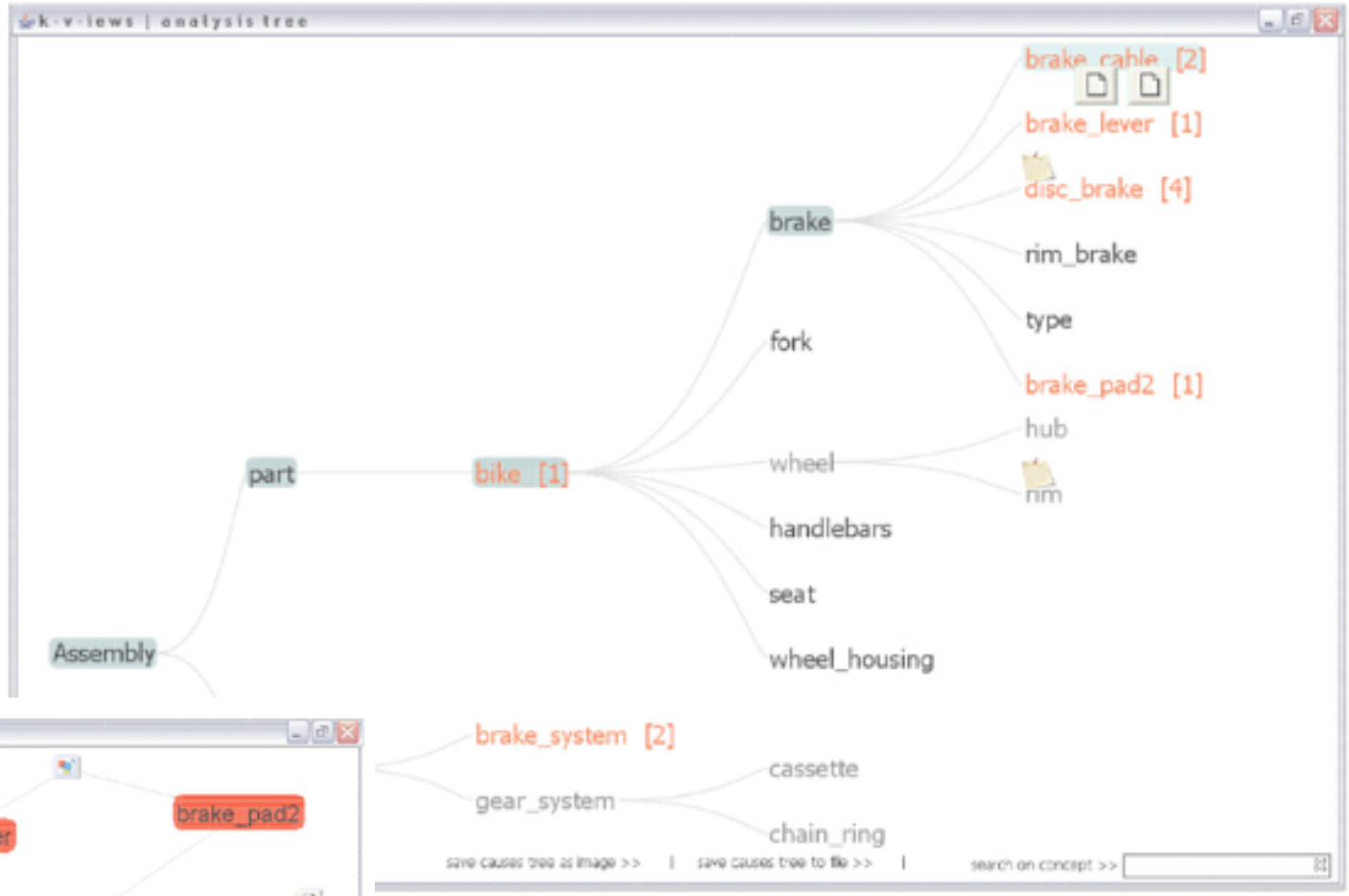
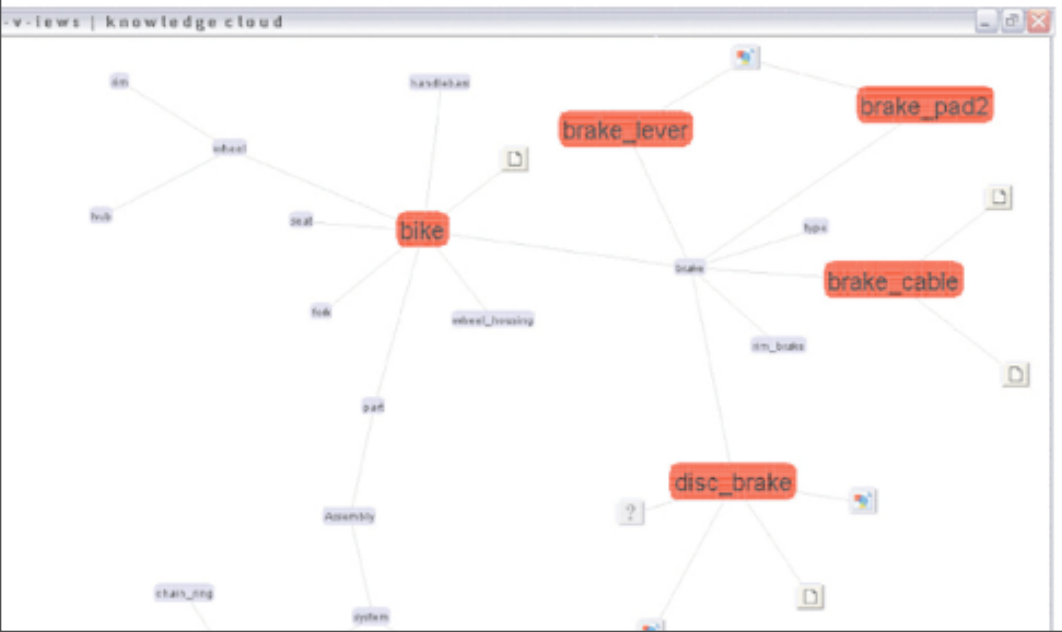


image cluster search
Ontology Browser
ontology
perspectives
Search
Graph
Search Criteria
No. of Matches
Image Size
small
Show Selected Only
Results
Image
Label
calliper brakes
part type = calliper
content-based image retrieval
Results

Knowledge Lenses





The University Of Sheffield.

Conclusions

About Semantic KM

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

On KM and SW

- The Semantic WEB offers potentially key technologies to the development of future knowledge Management and the Web
  - More Web than Semantics, but:
    - A little semantics goes a long way (J. Hendler)
- The potential must be exploited addressing real world requirements
  - Rather than in principle AI-oriented requirements (e.g. closed world, small scale, etc.)
- Strong application pull can be obtained
  - Do not sell slogans, sell ideas and applications!
- Recent industrial funders of my activity:

Rolls-Royce Kodak Environment Agency LYCOS

- Our spin-out company k-now.co.uk

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

A final thought

- These technologies allow easy collection of and access to a *very* large amount of information/knowledge
- Are we:
  - Preparing for a better Web/better world?
  - Preparing for a world with no privacy?
    - Big brother
    - Spam
    - Identity theft (e.g. Garlik)
  - Just adding hay to the haystack while searching for a needle?
    - Drowning in triples while trying to avoid drowning in texts?

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Bibliography

© Fabio Ciravegna, University of Sheffield

Semantic Search: biblio

- Uren, V., Lei, Y., Lopez, V., Liu, H., Motta, E.and Giordanino, M.: The usability of semantic search tools: a review, Knowledge Engineering Review, in press.
- Kaufmann, E. and Bernstein, A.: How Useful are Natural Language Interfaces to the Semantic Web for Casual End-users? Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, Busan, Korea, November 2007
- Lei, Y., Uren, V. and Motta, E. SemSearch: A Search Engine for the Semantic Web. in 15th International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks (EKAW 2006). 2006. Podebrady.
- Guha, R., McCool, R. Miller, E. Semantic Search. in 12th International Conference on World Wide Web. 2003
- Gilardoni, L., Biasuzzi, C., Ferraro, M., Fonti, R., Slavazza, P.: LKMS – A Legal Knowledge Management System exploiting Semantic Web technologies, Proceedings of the 4th International Conference on the Semantic Web (ISWC), Galway, November 2005.
- Rocha, R., Schwabe, D. and Poggi de Aragão, M.: A Hybrid Approach for Searching in the Semantic Web, in the 2004 International World Wide Web Conference, May 17-22, 2004, New York, New York.
- Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Vitaveska Lanfranchi and Daniela Petrelli: Hybrid Search: Effectively Combining Keywords and Semantic Searches in Proceedings of the 5th European Semantic Web Conference, ESWC 08, Tenerife, June 2008
- Tran, T., Cimiano, P., Rudolph, R. and Studer, R.: Ontology-based Interpretation of Keywords for Semantic Search. Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, Busan, Korea, November 2007
- Catarci, T., Di Mascio, T., Franconi, E., Santucci, G., Tessaris, S. An Ontology Based Visual Tool for Query Formulation Support. in 16th European Conference on Artificial Intelligence (ECAI-04). 2004. Valencia, Spain.
- Kaufmann, E., Bernstein, A. and Zumstein, R. Querix: A natural language interface to query ontologies based on clarification dialogs. In 5th ISWC, pages 980–981, Athens, GA, 2006.
- Corby, O., Dieng-Kuntz, R., Faron-Zucker, C., and Gandon, F., Searching the Semantic Web: Approximate Query Processing Based on Ontologies. IEEE Intelligent Systems, 2006. 21(1)

© Fabio Ciravegna, University of Sheffield

Automatic Document Annotation: bib

- Fabio Ciravegna. Designing adaptive information extraction for the Semantic Web in Amilcare. In S. Handschuh and S. Staab, editors, Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications. IOS Press, 2003.
- Fabio Ciravegna, Sam Chapman, Alexiei Dingli, and Yorick Wilks: Learning to Harvest Information for the Semantic Web, Proceedings of the First European Semantic Web Conference, Crete, May 2004
- A. Kiryakov, B. Popov, et al. Semantic Annotation, Indexing, and Retrieval. 2nd International Semantic Web Conference (ISWC2003), http://www.ontotext.com/publications/index.html#KiryakovEtAl2003
- S. Dill, N. Eiron, et al: http://www.tomkinshome.com/papers/2Web/semtag.pdf . SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. WWW'03.
- Thomas Leonard and Hugh Glaser. Large scale acquisition and maintenance from the web without source access. In Siegfried Handschuh, Rose Dieng-Kuntz, and Steffen Staab, editors, Proceedings Workshop 4, Knowledge Markup and Semantic Annotation, K-CAP 2001, 2001
- Ireson, N., Ciravegna, F., Califf, M.E., Freitag, D., Kushmerick, N., Lavelli, A.: Evaluating Machine Learning for Information Extraction, Proceedings of the 22nd International Conference on Machine Learning (ICML 2005), Bonn, Germany, 2005
- Iria, J. and Ciravegna, F A Methodology and Tool for Representing Language Resources for Information Extraction. In Proc. of LREC 2006, Genoa, Italy, May 2006.
- F. Ciravegna: Challenges in Information Extraction from Text for Knowledge Management, in S. Staab, (ed), "Human Language Technologies for Knowledge Management", IEEE Intelligent Systems and Their Applications (Trends and Controversies), Vol. 16, No. 6, pp 88-90, 2001.
- Fabio Ciravegna. Adaptive information extraction from text by rule induction and generalisation. In Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI), 2001. Seattle.
- H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). 2002.
- I. Muslea, S. Minton, and C. Knoblock. 1998. Wrapper induction for semistructured webbased information sources. In Proceedings of the Conference on Automated Learning and Discovery (CONALD), 1998.
- Chakravarthy, A., Lanfranchi, V., Ciravegna, F.: Cross-media Document Annotation and Enrichment, Proceedings of the 1st Semantic Authoring and Annotation Workshop, 5th International Semantic Web Conference (ISWC2006), Athens, GA, USA, 2006
- Handschuh, Staab, Ciravegna. S-CREAM - Semi-automatic CREAtion of Metadata (2002) http://citeseer.nj.nec.com/529793.html

© Fabio Ciravegna, University of Sheffield

Sharing, Reuse and Visualisation: biblio

- Dzbor, M. - Domingue, J. B. - Motta, E.: Magpie - towards a semantic web browser. 2nd International Semantic Web Conference (ISWC), Sanibel Island, Florida, USA, 2003.
- Lanfranchi, V., Ciravegna, F., Petrelli, D.: Semantic Web-based Document: Editing and Browsing in AktiveDoc, Proceedings of the 2nd European Semantic Web Conference , Heraklion, Greece, 2005.
- Fluit, C., Sabou, M., van Harmelen, F. Ontology-based Information Visualization. In: Geroimenko, V. and Chen C. (eds.) Springer (2003).
- Shneiderman, B. The eye have it: A task by data type taxonomy of information visualization. In Bederson, B. and Shneiderman, B. The craft of information visualization. Morgan Kaufman (2003)
- Paul Mutton and Jennifer Golbeck: Visualization of Semantic Metadata and Ontologies", in Proceedings of the Seventh International Conference on Information Visualization, 2003
- Stuckenschmidt, H. et al. Exploring Large Document RepositThe DOPE Project, IEEE Computer Society, May/June (2004).
- schraefel, m.c., Wilson, M., Russell, A., Smith, D. mSpace: Improving Information Access to Multimedia Domains with MultiModal Exploratory Search. CACM, 49 (4) 47-49 (2006)
- Fluit, C., Sabou, M., van Harmelen, F. Ontology-based Information Visualization. In: Geroimenko, V. and Chen C. (eds.) Springer (2003).
- Geroimenko, V., Chen, C. (eds.) Visualizing the Semantic Web, Springer, (2003)
- Mutton, P., Golbeck, J. Visualization of Semantic Metadata and Ontologies. 7th International Conference on Information Visualization, IEEE Computer Society (2003)

© Fabio Ciravegna, University of Sheffield