1

image ® Rolls-Royce

Semantic Web Technologies for Capturing, Sharing and Reusing Knowledge

- aka Annotating and Searching the Semantic Web

- aka: HLT and ML for the SW

Sponsored by X-MEDIA
www.x-media-project.org

Professor Fabio Ciravegna
Web Intelligence Technology Lab,
Department of Computer Science,
University of Sheffield
http://www.dcs.shef.ac.uk/~fabio/
fabio@dcs.shef.ac.uk

SSSW-2008

Copyright Notice

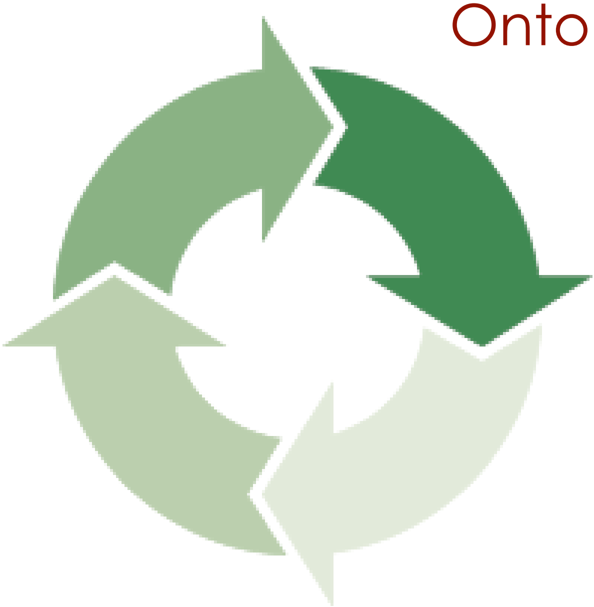
- These slides were presented during the The Sixth Summer School on Ontological Engineering and the Semantic Web (SSSW'08), July 6-12, 2008. Cercedilla(Spain) (http://kmi.open.ac.uk/events/sssw08/)

- Condition of use:
  - the use is limited to personal or educational purposes
  - the copyright footnote must always be visible when slides are presented
  - full recognition is given to me for the paternity of the slides and 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.

- Fabio Ciravegna, University of Sheffield,
  fabio@dcs.shef.ac.uk
  http://www.dcs.shef.ac.uk/~fabio/

The Knowledge Life-cycle

Acquisition and Modelling
Ontologies



Capture
Annotations

Sharing

Reuse
Knowledge

Note the gap!!!
Annotations are
not knowledge!!!!!

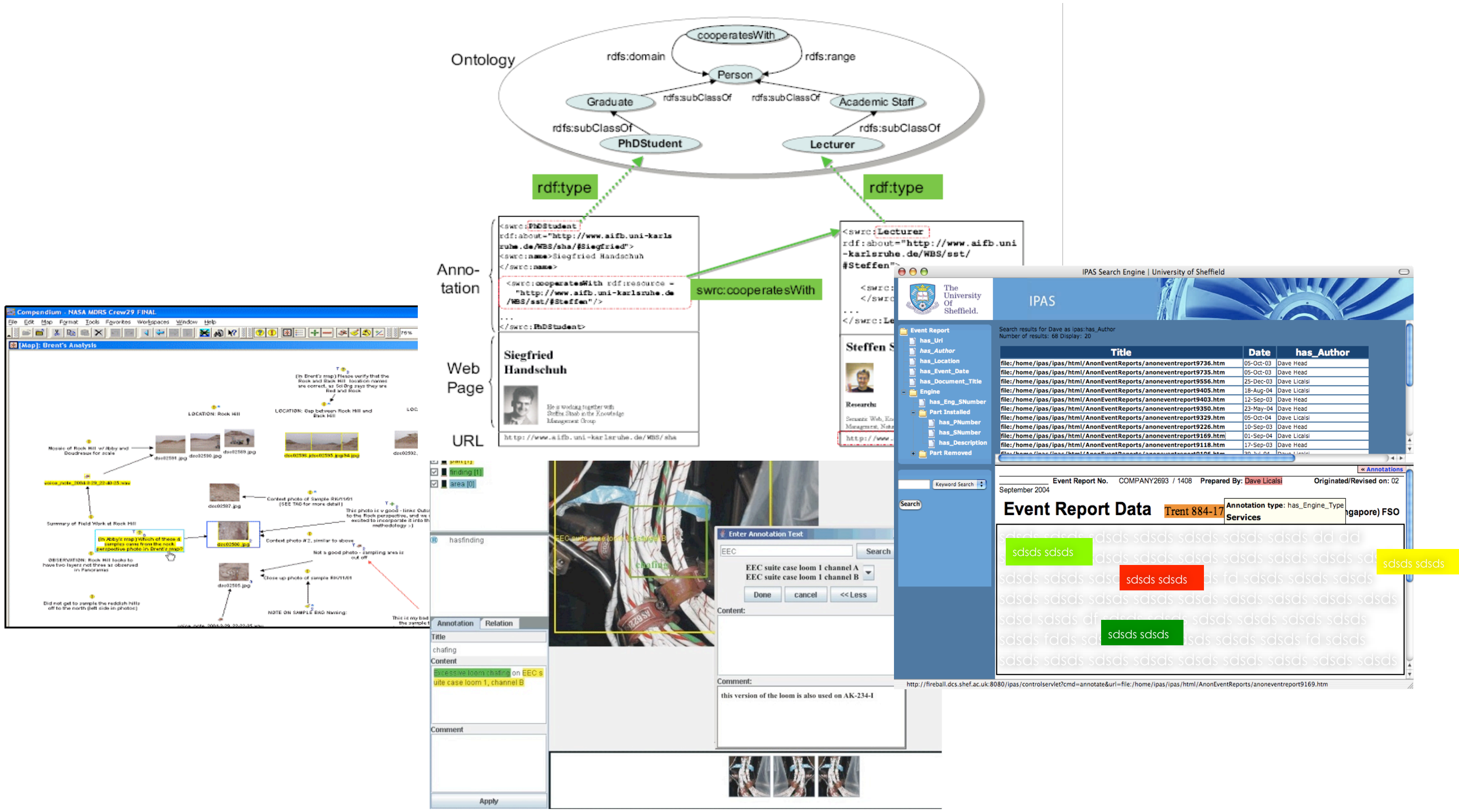
Why manage knowledge?

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

Please note: most books and tutorial work the other way around.
They start with modelling (e.g. ontology building) then move to acquisition, then to sharing (if they do!). This often generates confusion: modelling seems the most important issue!!

Today’s tutorial

- We will see techniques and methodologies for
  - Knowledge Capture
    - Extracting and integrating information
      - from existing archives and documents
      - With user in the loop
  - Knowledge Sharing and Reuse
    - Enabling knowledge searching + process support
- You have already seen:
  - Knowledge Acquisition and Modelling
    - Ontology Engineering



Requirements for Knowledge Capture

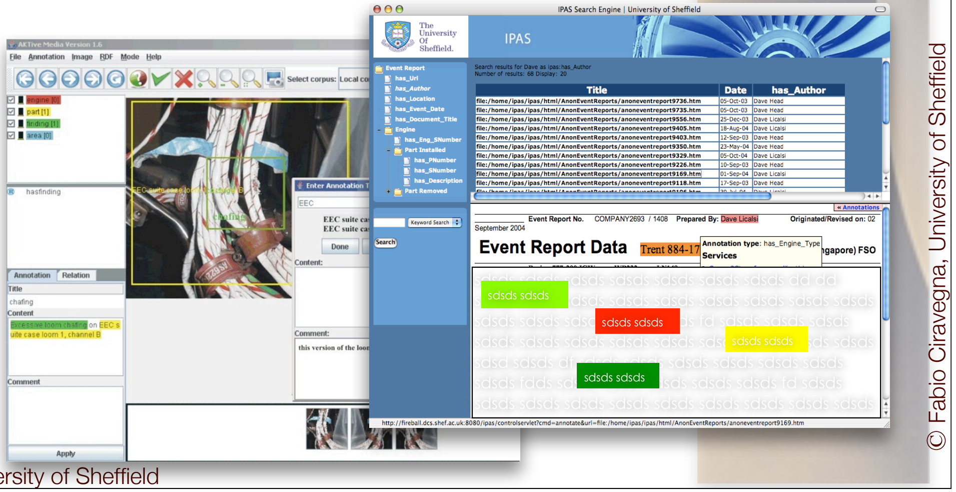
- issues in knowledge capture:
  - capture: what and what for?

Knowledge Capture

- Collecting and aggregating multimedia knowledge to make it available for
  - sharing and reuse
    - From document management to knowledge management
  - for integration

- Approaches
  - at source: helping people capturing knowledge when produced
- On legacy documents, pictures, data:
  - Annotation services

In ontological terms knowledge capture consists in capturing instances!



Title: chafing

Content: Excessive loom chafing on EEC suite case loom 1, channel B

EEC

EEC suite ca
EEC suite ca

Content:

Comment: this version of the loo

Search results for Dave as ipas:has_Author
Number of results: 68 Display: 20

Title	Date	has_Author
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9736.htm	05-Oct-03	Dave Head
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9735.htm	05-Oct-03	Dave Head
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9556.htm	25-Dec-03	Dave Licalsi
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9405.htm	18-Aug-04	Dave Licalsi
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9403.htm	12-Sep-03	Dave Head
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9350.htm	23-May-04	Dave Head
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9329.htm	05-Oct-04	Dave Licalsi
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9226.htm	10-Sep-03	Dave Head
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9169.htm	01-Sep-04	Dave Licalsi
file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9118.htm	17-Sep-03	Dave Head

Event Report No. COMPANY2693 / 1408 Prepared By: Dave Licalsi Originated/Revised on: 02 September 2004

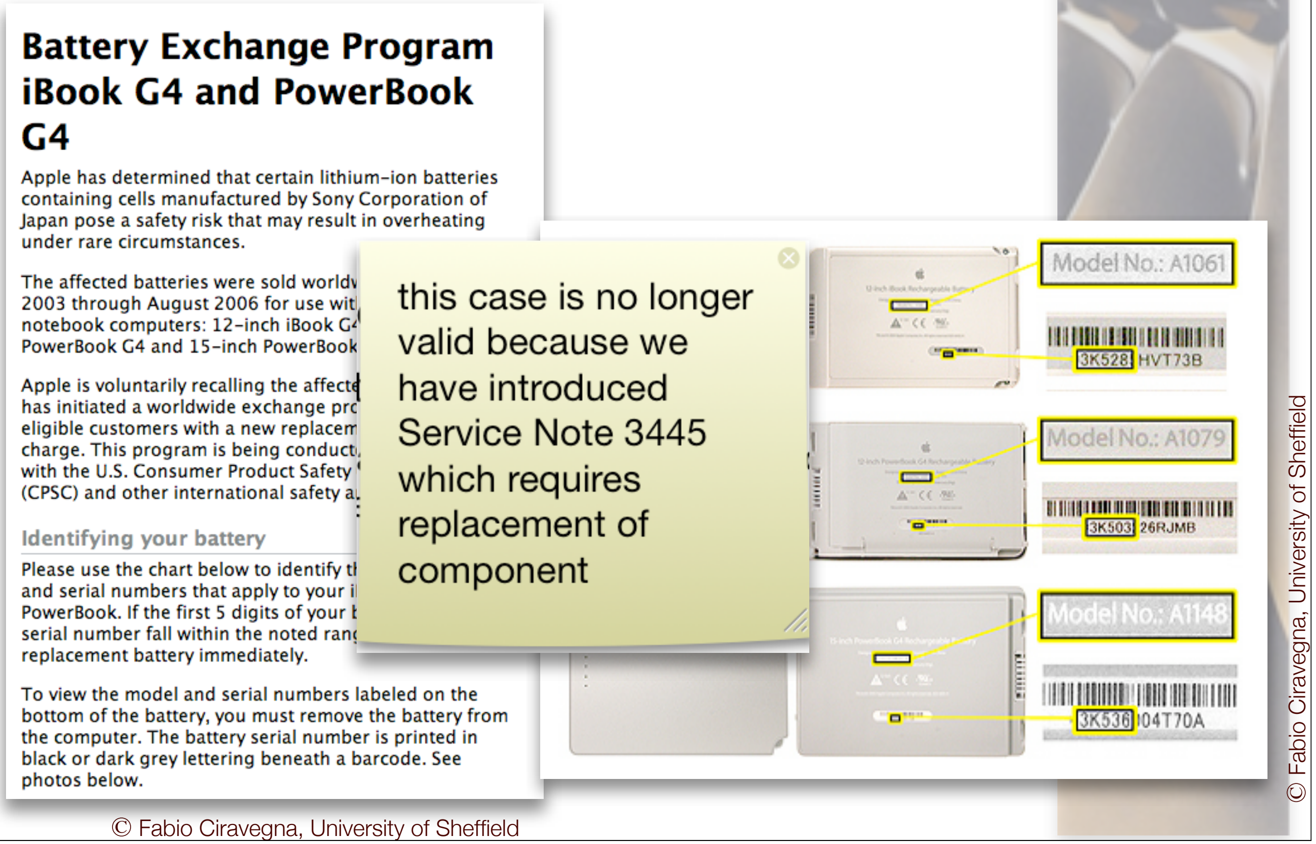
Event Report Data Trent 884-17 ngapore) FSO

Annotation type: has_Engine_Type
Services

http://fireball.dcs.shef.ac.uk:8080/ipas/controlservlet?cmd=annotate&url=file:/home/ipas/ipas/html/AnonEventReports/anoneventreport9169.htm

Requirements for KC: Cross media

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



Battery Exchange Program iBook G4 and PowerBook G4

Apple has determined that certain lithium-ion batteries containing cells manufactured by Sony Corporation of Japan pose a safety risk that may result in overheating under rare circumstances.

The affected batteries were sold worldv
2003 through August 2006 for use wit
notebook computers: 12-inch iBook G4
PowerBook G4 and 15-inch PowerBook

Apple is voluntarily recalling the affecte
has initiated a worldwide exchange pro
eligible customers with a new replacem
charge. This program is being conduct
with the U.S. Consumer Product Safety
(CPSC) and other international safety a

Identifying your battery

Please use the chart below to identify t
and serial numbers that apply to your i
PowerBook. If the first 5 digits of your b
serial number fall within the noted ran
replacement battery immediately.

To view the model and serial numbers labeled on the bottom of the battery, you must remove the battery from the computer. The battery serial number is printed in black or dark grey lettering beneath a barcode. See photos below.

this case is no longer valid because we have introduced Service Note 3445 which requires replacement of component

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 within the same physical document
    - Information contained in just one modality is insufficient
    - Cross-media knowledge acquisition techniques needed in order to capture and manage all of the explicit and implicit knowledge

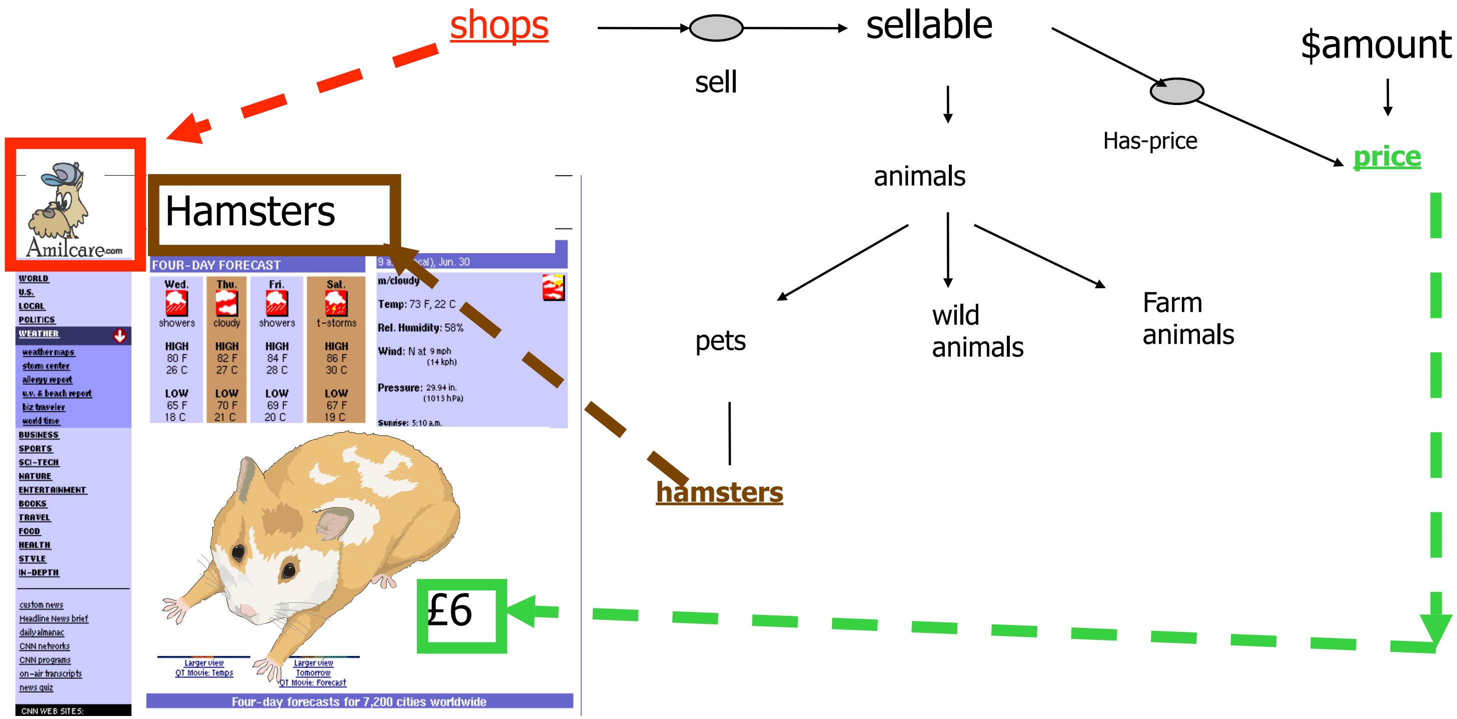
A way of thinking

The inside of every Yaris looks clean and sophisticated. The controls and instruments are ergonomically designed and positioned. And there is a distinct lack of clutter thanks to the innovative storage system.

Every inch of space has been used to provide a range of practical storage compartments that ensure items are neatly and safely stowed. The glove box, for example, has two sections, while a useful tray under the passenger seat keeps the contents out of sight.

There are two large storage pockets next to the centre console, two front-door storage bins, cup-holders and a passenger's side storage tray. Additionally, the T Spirit and T Sport models have pockets in the back of the front seats.

The Yaris – neatly taking care of everything.



FOUR-DAY FORECAST

Wed.	Thu.	Fri.	Sat.
showers	cloudy	showers	t-storms
HIGH 80 F 26 C	HIGH 82 F 27 C	HIGH 84 F 28 C	HIGH 86 F 30 C
LOW 65 F 18 C	LOW 70 F 21 C	LOW 69 F 20 C	LOW 67 F 19 C

9 a. cal), Jun. 30

m/cloudy

Temp: 73 F, 22 C

Rel. Humidity: 58%

Wind: N at 9 mph (14 kph)

Pressure: 29.94 in. (1013 hPa)

Sunrise: 5:10 a.m.

Four-day forecasts for 7,200 cities worldwide

SW for Knowledge Capture

- user centred methodologies and tools for text and image annotation
- automatic methodologies and tools for text annotation

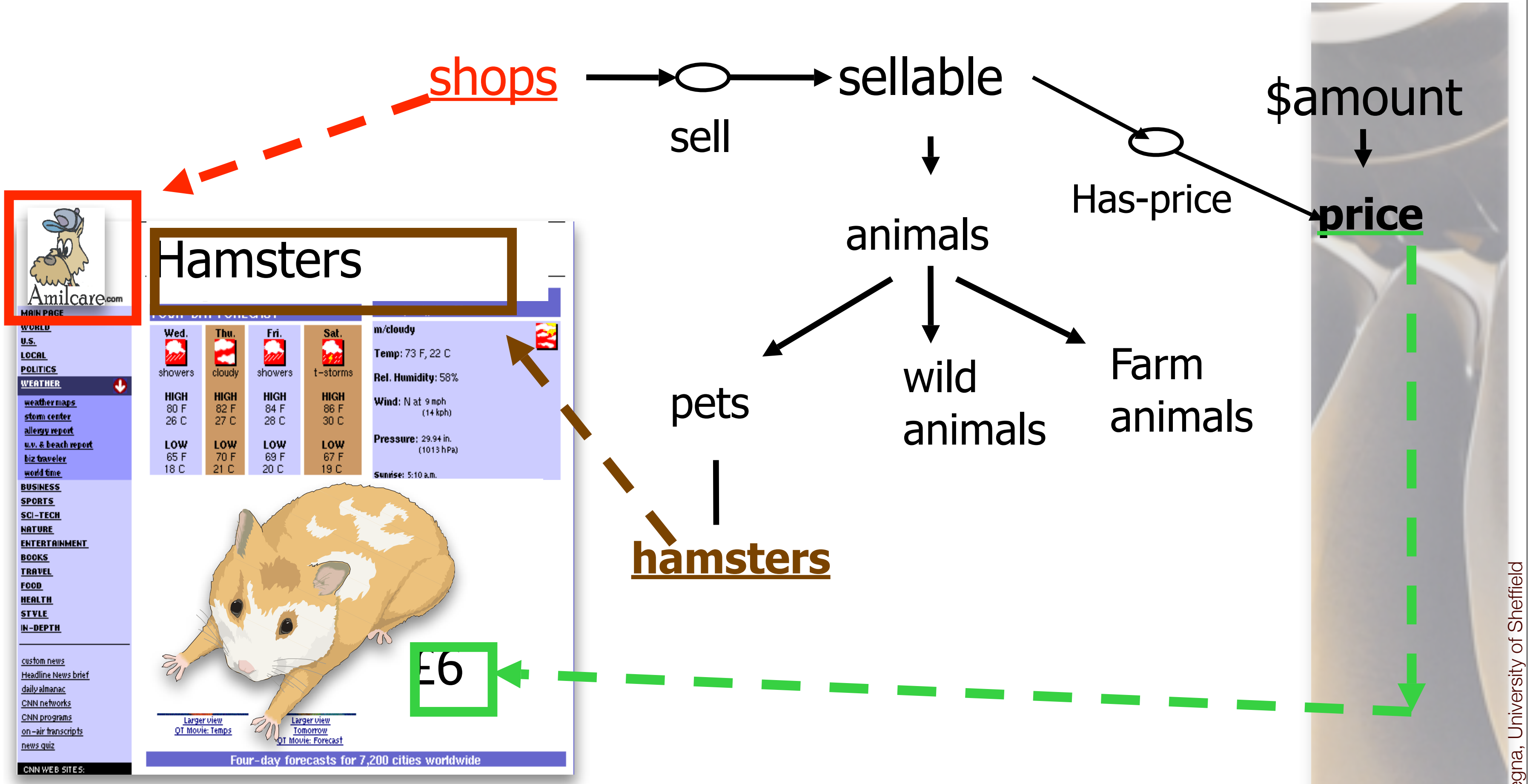
Semantic Web for Knowledge Capture

- Aims:
  - To capture knowledge within and across media in a rich, semantically-oriented way
  - Outcome of capture technologies is a semantic representation of the content (conceptualisation) to be used for knowledge management purposes
  - Enrichment of multimedia documents with layers of manually or automatically generated annotation is the main medium of associating conceptualisations to resources

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...

Ontology-based Annotation



shops

sell

sellable

Has-price

$amount

price

animals

pets

wild animals

Farm animals

hamsters

Amilcare.com

Hamsters

MAIN PAGE
WORLD
U.S.
LOCAL
POLITICS
WEATHER
weather maps
storm center
allergy report
u.v. & beach report
biz traveler
world time
BUSINESS
SPORTS
SCI-TECH
NATURE
ENTERTAINMENT
BOOKS
TRAVEL
FOOD
HEALTH
STYLE
IN-DEPTH

custom news
Headline News brief
daily almanac
CNN networks
CNN programs
on-air transcripts
news quiz

CNN WEB SITES:

Wed.	Thu.	Fri.	Sat.
showers	cloudy	showers	t-storms
HIGH 80 F 26 C	HIGH 82 F 27 C	HIGH 84 F 28 C	HIGH 86 F 30 C
LOW 65 F 18 C	LOW 70 F 21 C	LOW 69 F 20 C	LOW 67 F 19 C

m/cloudy

Temp: 73 F, 22 C

Rel. Humidity: 58%

Wind: N at 9 mph (14 kph)

Pressure: 29.94 in. (1013 hPa)

Sunrise: 5:10 a.m.

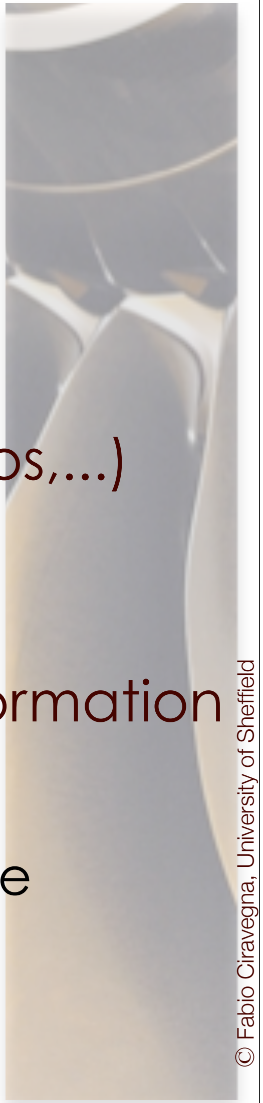
£6

Larger view
QT Movie: Temps

Larger view
Tomorrow
QT Movie: Forecast

Four-day forecasts for 7,200 cities worldwide

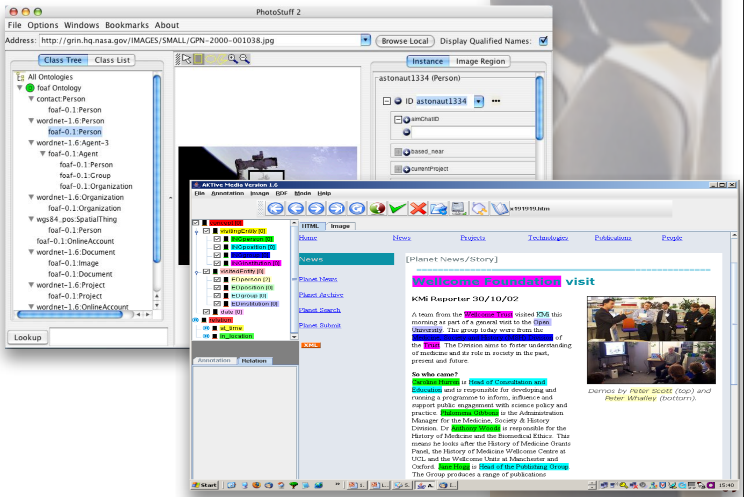
Input &Output



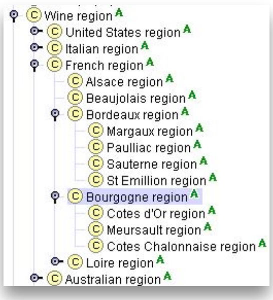
- Input to the KC technologies
  - Ontologies (MMO, domain ontology),
  - Background knowledge (gazetteers, etc.)
  - Normalised document representation
  - Medium to extract from (text, images, data, videos,...)
- Output
  - Evidence represented in terms of conceptual information
    - Evidence used by other modules as background conceptual knowledge, i.e. pre-existing knowledge
    - Evidence in the form of uncertain output

Ontology-based Annotation

- The way to annotate pages is to:
  - Select an ontology
  - Define statements to represent meta-data about the document
- Manual Annotation
  - Annotation can be performed by:
    - Domain expert
- User-friendly tools for annotation
  - Cream (Handschuh *et al.* 2002)
  - Melita (Ciravegna *et al.* 2002)
  - Photostuff (Hendler *et al.* 2005)
  - AktiveMedia (Chakravarthy *et al.* 2006)

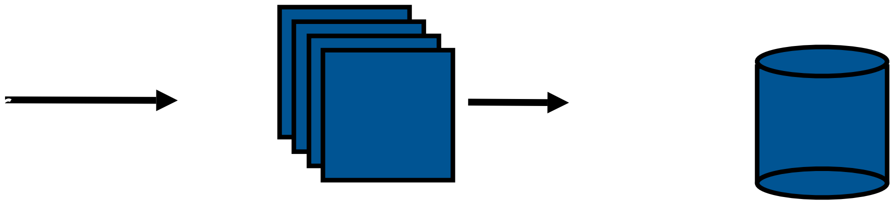


Ontology-based Annotation



- Wine region
  - United States region
  - Italian region
  - French region
    - Alsace region
    - Beaujolais region
    - Bordeaux region
      - Margaux region
      - Paulliac region
      - Sauterne region
      - St Emillion region
    - Bourgogne region
      - Cotes d'Or region
      - Meursault region
      - Cotes Chalonnaise region
    - Loire region
  - Australian region

Ontology



Annotated Documents

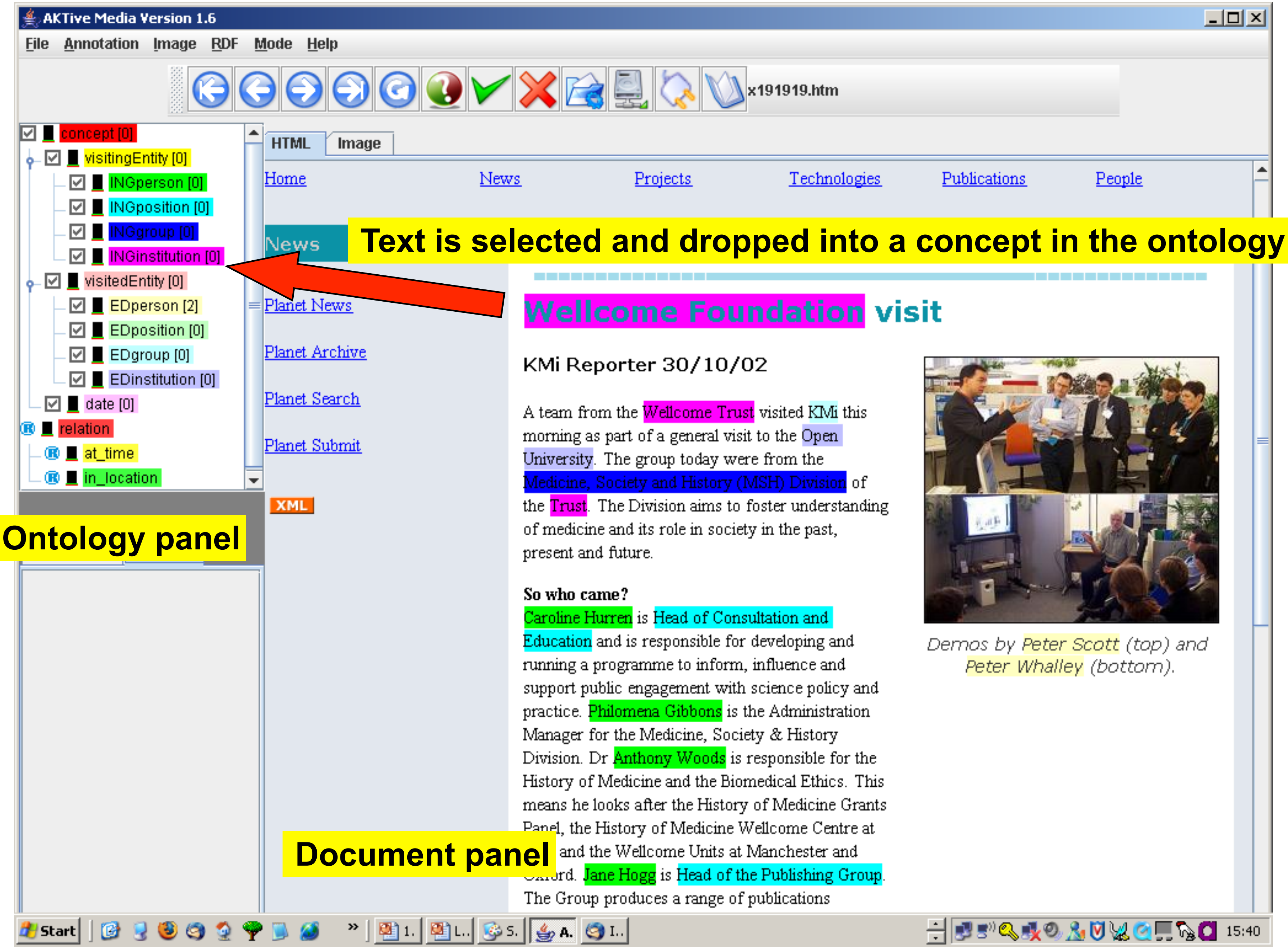
Triple store (annotations)

3store
Sesame
...

AktiveMedia

- Enables semi-automatic annotation across texts and images
- The interface enables
  - HTML editing
  - Annotation of documents in RDF based on an OWL ontology
- Types of annotations
  - Concepts / Relations
- SW: Annotation:
  - Selection of concept/relation and highlighting of text is the way in which annotation is performed

SS http://www.dcs.shef.ac.uk/~ajay/html/cresearch.html



Text is selected and dropped into a concept in the ontology

Ontology panel

Document panel

Wellcome Foundation visit

KMi Reporter 30/10/02

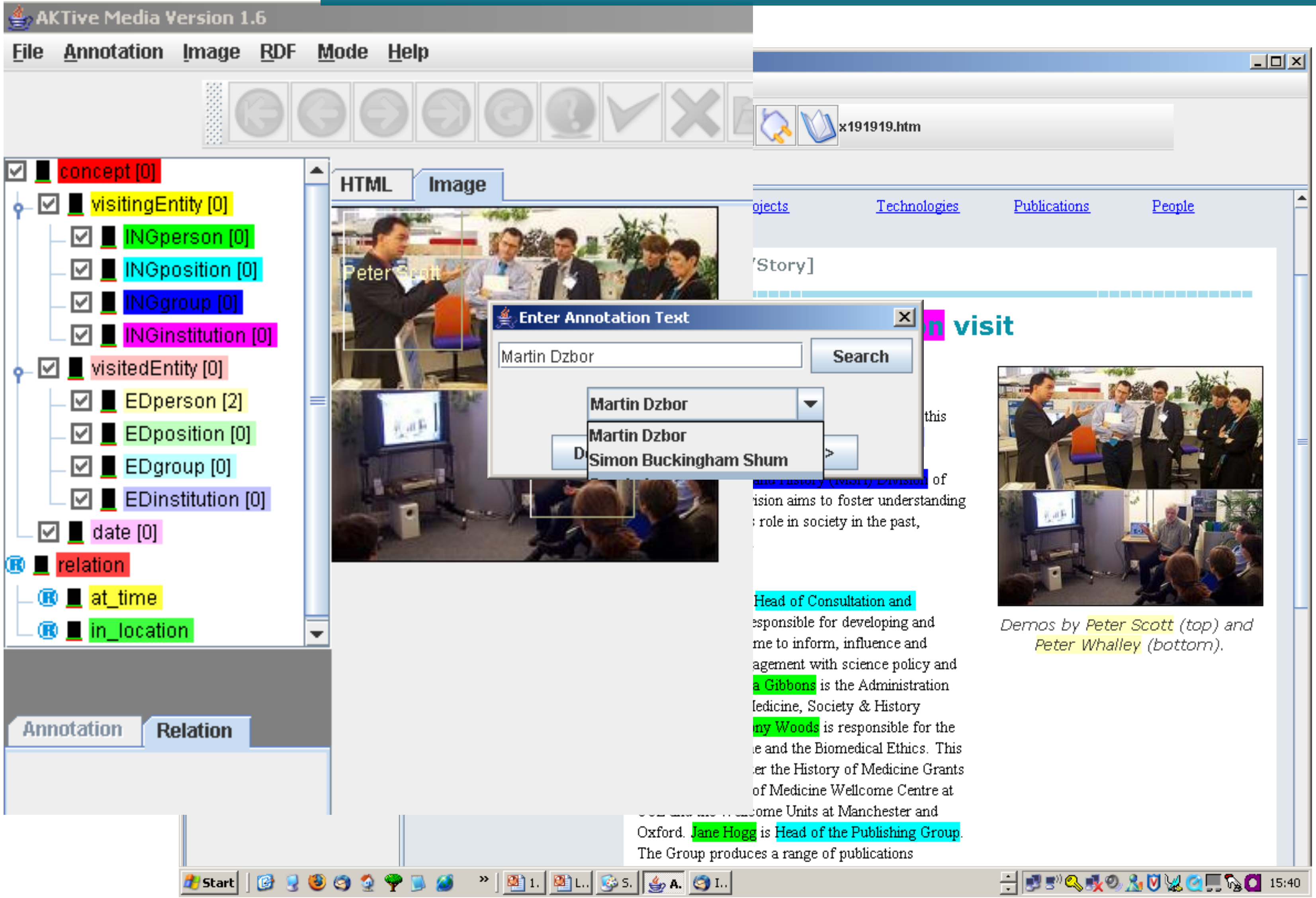
A team from the Wellcome Trust visited KMi this morning as part of a general visit to the Open University. The group today were from the Medicine, Society and History (MSH) Division of the Trust. The Division aims to foster understanding of medicine and its role in society in the past, present and future.

So who came?

Caroline Hurren is Head of Consultation and Education and is responsible for developing and running a programme to inform, influence and support public engagement with science policy and practice. Philomena Gibbons is the Administration Manager for the Medicine, Society & History Division. Dr Anthony Woods is responsible for the History of Medicine and the Biomedical Ethics. This means he looks after the History of Medicine Grants Panel, the History of Medicine Wellcome Centre at and the Wellcome Units at Manchester and Oxford. Jane Hogg is Head of the Publishing Group. The Group produces a range of publications

Demos by Peter Scott (top) and Peter Whalley (bottom).

Contextual Annotation of Images and Text



Enter Annotation Text: Martin Dzbor

Martin Dzbor

Martin Dzbor
Simon Buckingham Shum

'Story]

n visit

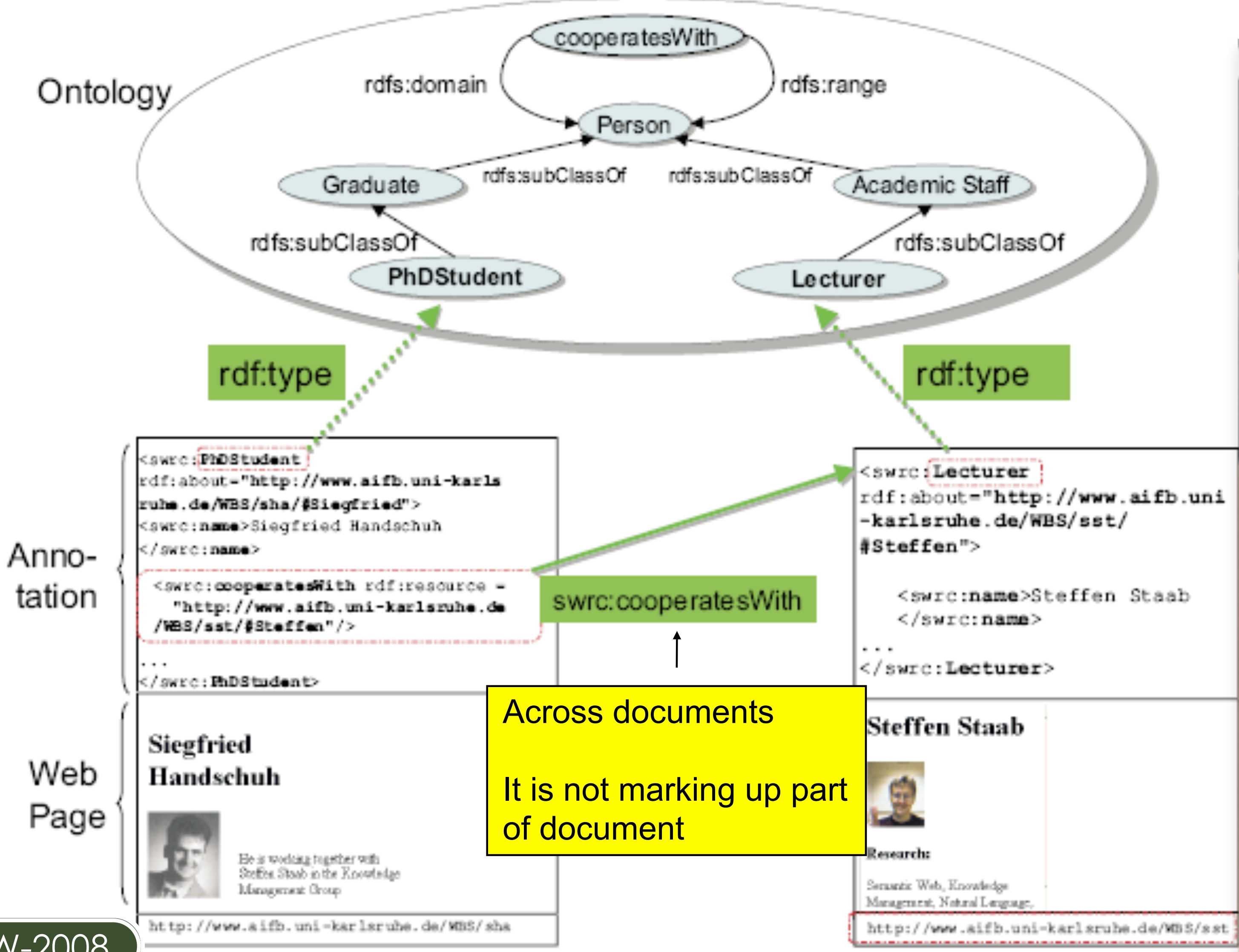
this

and History (MSH) Division of
ision aims to foster understanding
role in society in the past,

Head of Consultation and
esponsible for developing and
me to inform, influence and
agement with science policy and
a Gibbons is the Administration
ledicine, Society & History
ny Woods is responsible for the
e and the Biomedical Ethics. This
er the History of Medicine Grants
of Medicine Wellcome Centre at
UCL and the Wellcome Units at Manchester and
Oxford. Jane Hogg is Head of the Publishing Group.
The Group produces a range of publications

Demos by Peter Scott (top) and Peter Whalley (bottom).

Annotating across documents (CREAM, 2001)



```
<swrc:PhDStudent
rdf:about="http://www.aifb.uni-karls
ruhe.de/WBS/sha/#Siegfried">
<swrc:name>Siegfried Handschuh
</swrc:name>
 <swrc:cooperatesWith rdf:resource =
   "http://www.aifb.uni-karlsruhe.de
 /WBS/sst/#Steffen"/>
...
</swrc:PhDStudent>
```

```
<swrc:Lecturer
rdf:about="http://www.aifb.uni
-karlsruhe.de/WBS/sst/
#Steffen">

    <swrc:name>Steffen Staab
    </swrc:name>
...
</swrc:Lecturer>
```

Siegfried Handschuh

He is working together with Steffen Staab in the Knowledge Management Group

http://www.aifb.uni-karlsruhe.de/WBS/sha

Steffen Staab

Research:

Semantic Web, Knowledge Management, Natural Language,

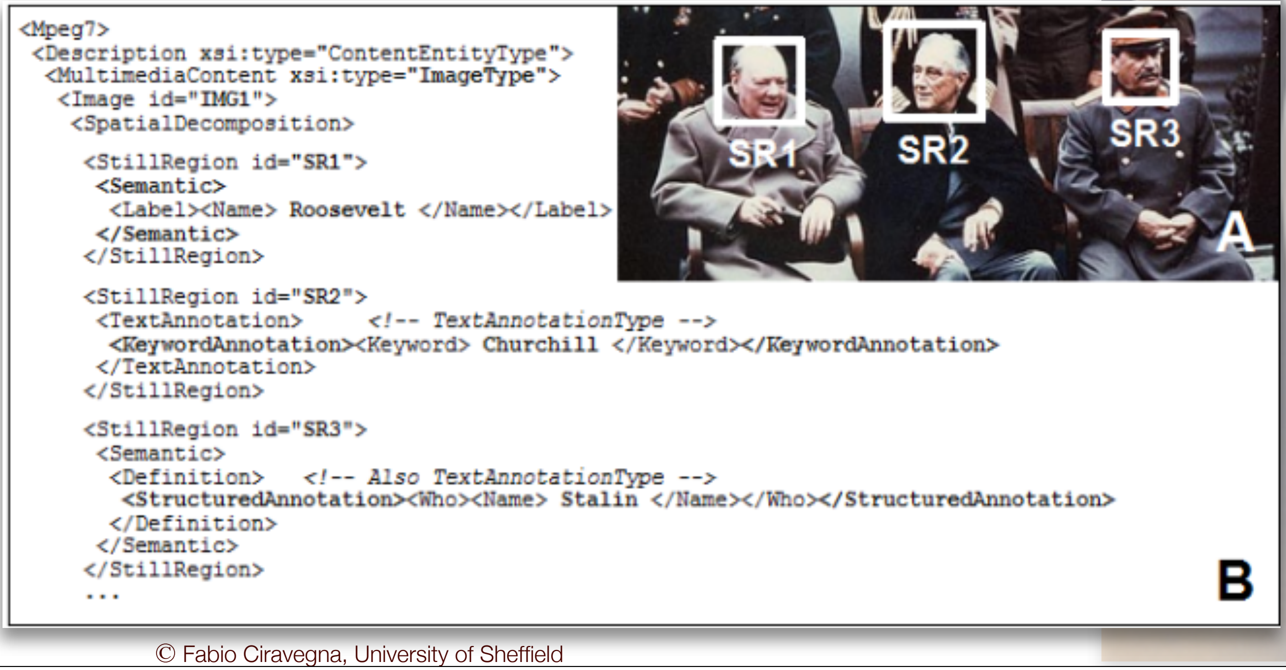
http://www.aifb.uni-karlsruhe.de/WBS/sst

Across documents

It is not marking up part of document

Marking up Provenance

- COMM - A Core Ontology for Multimedia based on http://comm.semanticweb.org/
  - the MPEG-7 standard
  - the DOLCE foundational ontology.

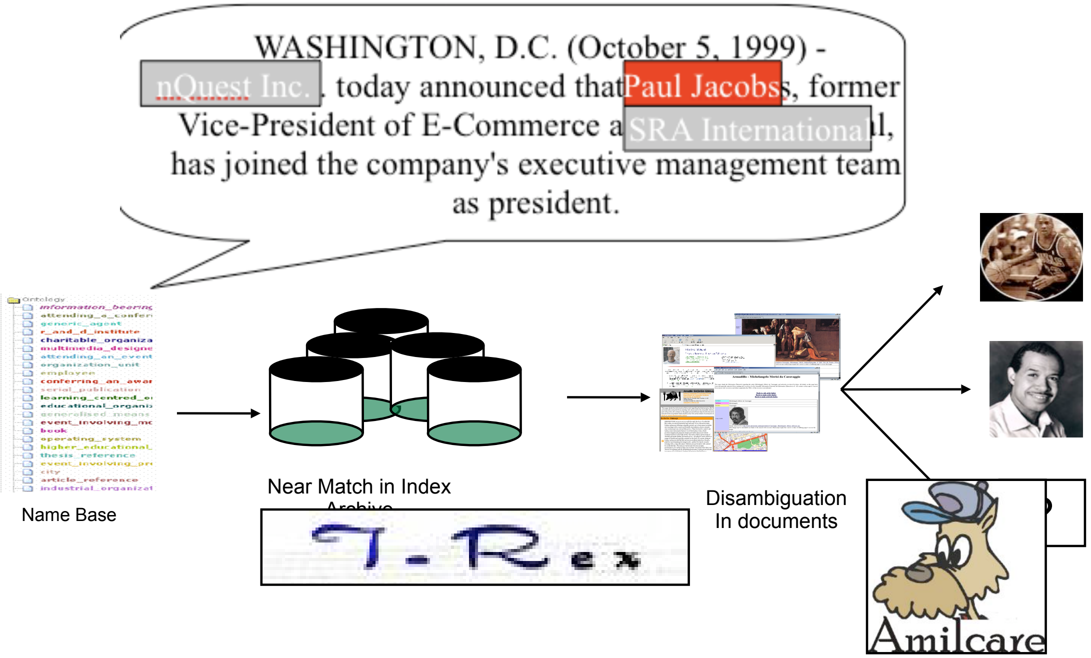


```
<Mpeg7>
 <Description xsi:type="ContentEntityType">
  <MultimediaContent xsi:type="ImageType">
   <Image id="IMG1">
    <SpatialDecomposition>

     <StillRegion id="SR1">
      <Semantic>
       <Label><Name> Roosevelt </Name></Label>
      </Semantic>
     </StillRegion>

     <StillRegion id="SR2">
      <TextAnnotation>     <!-- TextAnnotationType -->
       <KeywordAnnotation><Keyword> Churchill </Keyword></KeywordAnnotation>
      </TextAnnotation>
     </StillRegion>

     <StillRegion id="SR3">
      <Semantic>
       <Definition>   <!-- Also TextAnnotationType -->
        <StructuredAnnotation><Who><Name> Stalin </Name></Who></StructuredAnnotation>
       </Definition>
      </Semantic>
     </StillRegion>
     ...
```



WASHINGTON, D.C. (October 5, 1999) - nQuest Inc. today announced that Paul Jacobss, former Vice-President of E-Commerce a SRA Internationall, has joined the company's executive management team as president.

Automating Annotation

© Fabio Ciravegna, University of Sheffield

Annotation Engines

- Solutions like AktiveMedia can be used for annotating new documents and 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
    - ...

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

Information Extraction from Text

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

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,

Named Entities

Event Recognition

Growing complexity

© Fabio Ciravegna, Univers

Classic Tasks

- Information Extraction from Text:
  - Entity Extraction
  - Fields Extraction
  - Relation Extraction
  - Event Extraction
- Other (non Semantic) Tasks
  - Document Similarity
  - Text Categorization

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

Traditional approach to NER&C

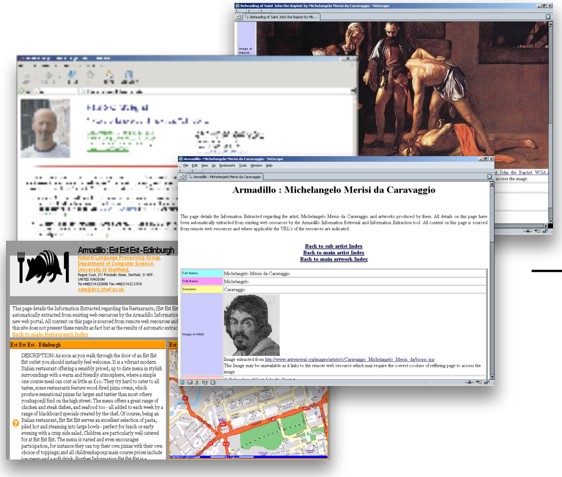
- Two steps:
  - Training phase
    - Input: annotated set of representative documents
    - Output: trained system
  - At runtime
    - One-by-one document analysis
- Expected accuracy:
  - 80-95% (free texts)
  - Web documents tend to require additional processing to get equivalent results (but doable to some extent)
- Medium Scale: up to hundreds of thousands of documents

Large Scale NER&C

- For large scale (some hundred millions pages) smarter infrastructure is needed
  - Search engine-like indexing infrastructure
  - Faster processing (less processing)
  - Two cases:
    - Recognition of known terms (and their variations)
      - See also information integration
    - Discovery of new names

Large Scale NER: Indexing

- Document Indexing as in Search Engines



Armadillo : Est Est Est - Edinburgh

Natural Language Processing Group,
Department of Computer Science,
University of Sheffield,
Regent Court, 211 Portobello Street, Sheffield, S1 4DP,
UNITED KINGDOM
Tel:+44(0)114-2228000 Fax:+44(0)114-22.21810
sam@dcs.shef.ac.uk

This page details the Information Extracted regarding the Restaurants, (Est Est Est automatically extracted from existing web resources by the Armadillo Information new web portal. All content on this page is sourced from remote web resources and this site does not present these results as fact but as the results of automatic extrac

Back to main Restaurants Index

Est Est Est - Edinburgh

DESCRIPTION: As soon as you walk through the door of an Est Est Est outlet you should instantly feel welcome. It is a vibrant modern Italian restaurant offering a sensibly priced, up to date menu in stylish surroundings with a warm and friendly atmosphere, where a simple one course meal can cost as little as £10. They try hard to cater to all tastes, some restaurants feature wood-fired pizza ovens, which produce sensational pizzas far larger and tastier than most others you&apos;ll find on the high street. The menu offers a great range of chicken and steak dishes, and seafood too - all added to each week by a range of blackboard specials created by the chef. Of course, being an Italian restaurant, Est Est Est serves an excellent selection of pasta, piled hot and steaming into large bowls - perfect for lunch or early evening with a crisp side salad. Children are particularly well catered for at Est Est Est. The menu is varied and even encourages participation, for instance they can top their own pizzas with their own choice of toppings; and all children&apos;s main course prices include

Armadillo : Michelangelo Merisi da Caravaggio - Netscape

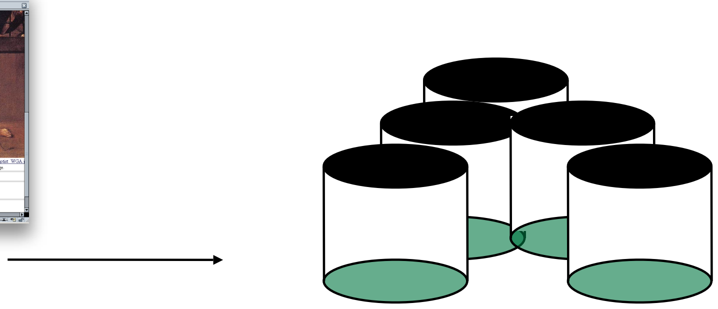
Armadillo : Michelangelo Merisi da Caravaggio

This page details the Information Extracted regarding the artist, Michelangelo Merisi da Caravaggio and artworks produced by them. All details on this page have been automatically extracted from existing web resources by the Armadillo Information Retrevial and Information Extraction tool. All content on this page is sourced from remote web resources and where applicable the URL's of the resources are indicated.

Back to sub artist Index
Back to main artist Index
Back to main artwork Index

Full Name:	Michelangelo Merisi da Caravaggio
Fist Name:	Michelangelo
Surname:	Caravaggio
Image of Artist:	Image extracted from http://www.artrenewal.org/images/artists/c/Caravaggio_Michelangelo_Merisi_da/biopic.jpg This Image may be unavailable as it links to the remote web resource which may require the correct cookies of reffering page to access the image.

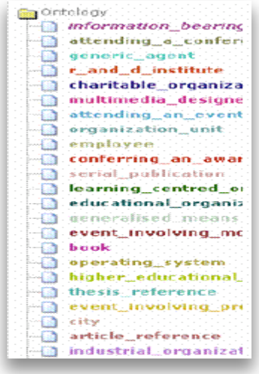
access the image.



Distributed Index Archive
(keywords)

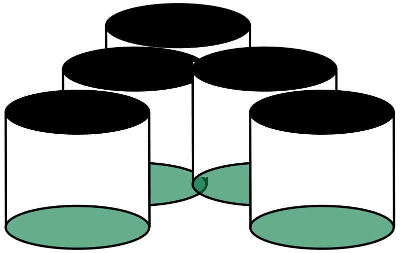
© Fabio Ciravegna, University of Sheffield

Known Name Recognition



Ontology
- information_bearing
- attending_a_confer
- generic_agent
- r_and_d_institute
- charitable_organiza
- multimedia_designe
- attending_an_event
- organization_unit
- employee
- conferring_an_awar
- serial_publication
- learning_centred_o
- educational_organi
- generalised_means
- event_involving_mc
- book
- operating_system
- higher_educational_
- thesis_reference
- event_involving_pr
- city
- article_reference
- industrial_organizat

Name Base



Near Match in Index Archive

Disambiguation
In documents

?

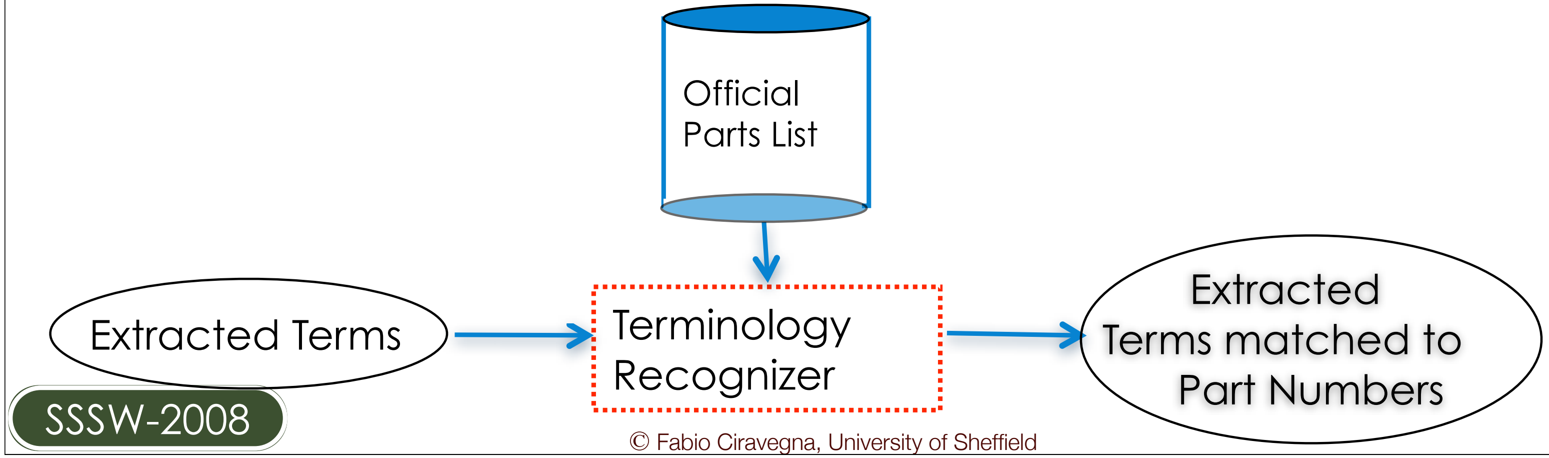
S. Dill, N. Eiron, et al: SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. WWW'03

Discovery of New Names

- Modified Indexing of documents to recognize potential names
  - Traditional NER
    - On the window of words (not the whole doc!!!)
      - Fast and effective
  - Web specific strategies
    - To identify names without context

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



Terminology Recognition

- Possible approaches
  - Linguistic approaches
    - Based on linguistic analysis of terms (Gaizauskas *et al* 2003)
  - Statistical approaches
    - Based on frequency analysis and detection
  - Other approaches
    - Distance metrics based (Butters 2007)

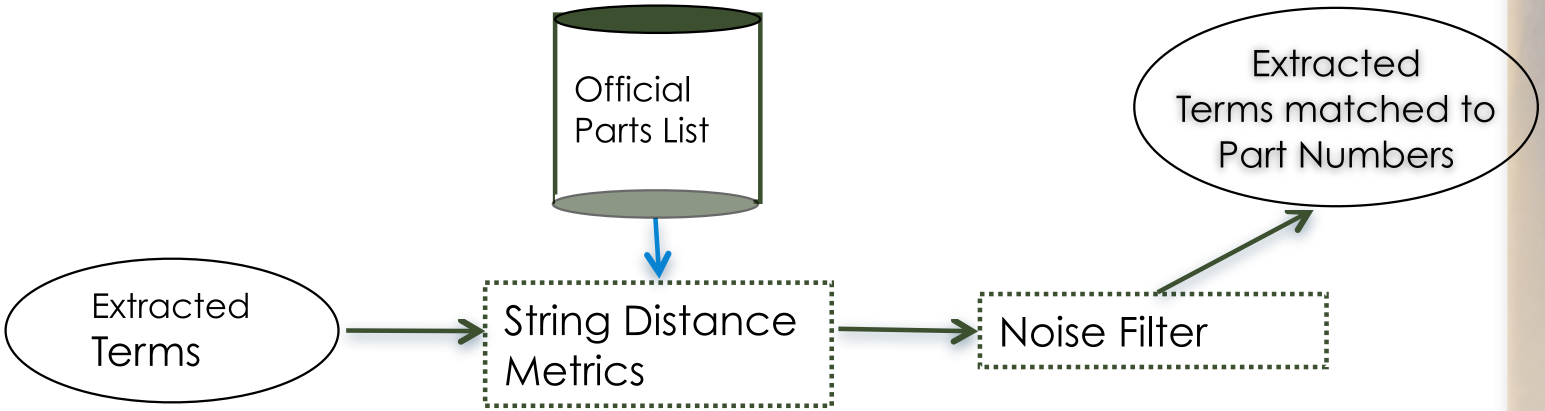


Table Field Extraction

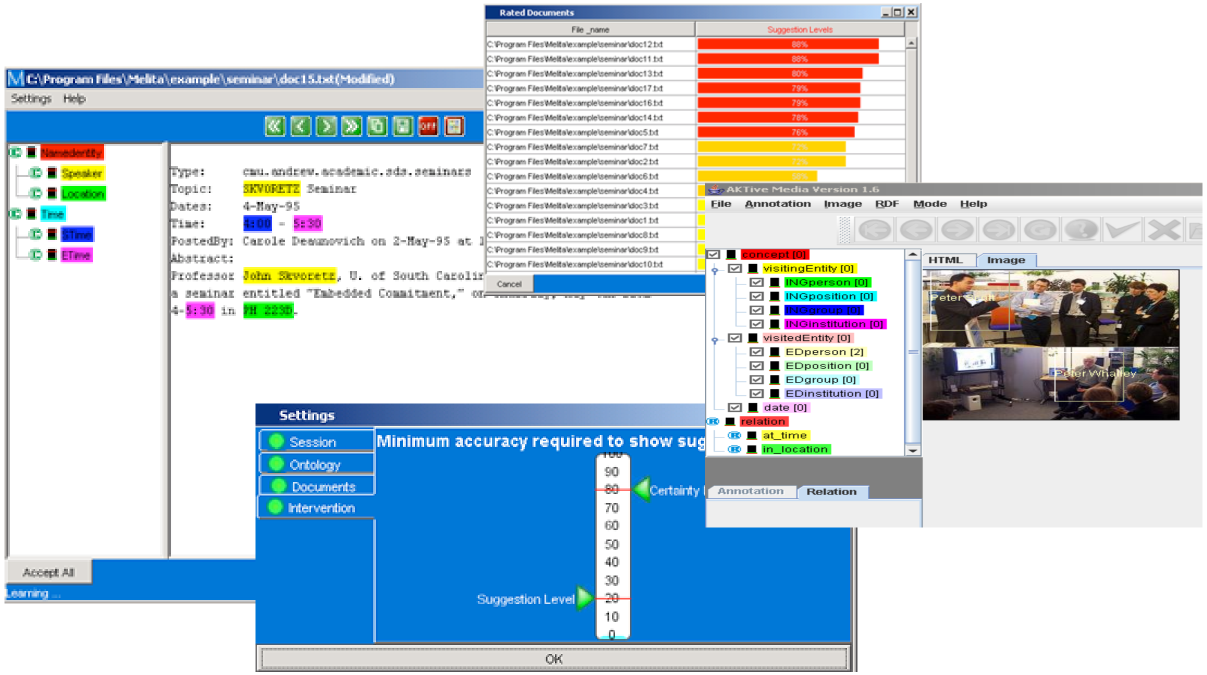
- Tables are an essential part of many documents
  - Most information is represented in tables
- Tables can be represented as forms to fill
  - Semantics is fixed
  - Wrapper writing or wrapper induction (Kushmerick 1997)
- Tables can be created ad hoc in documents (e.g. Word docs)
  - Semantics is unclear
  - 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

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

A list of tools for automatic annotation

- Architectures for IE:
  - UIMA (http://www.research.ibm.com/UIMA/)
  - GATE (www.gate.ac.uk)
    - Contains Annie: Named Entity Recogniser
  - KIM (http://www.ontotext.com/kim/)
- WiT toolbox: http://nlp.shef.ac.uk/wig/tools/)
  - Manual and semi-automatic annotation of texts and images
    - AktiveMedia http://www.dcs.shef.ac.uk/~ajay/html/cresearch.html
  - TRex: plugin for Machine Learning based IE http://tyne.shef.ac.uk/t-rex/index.html
  - Saxon: rule-based (FST) tool http://nlp.shef.ac.uk/wig/tools/saxon/

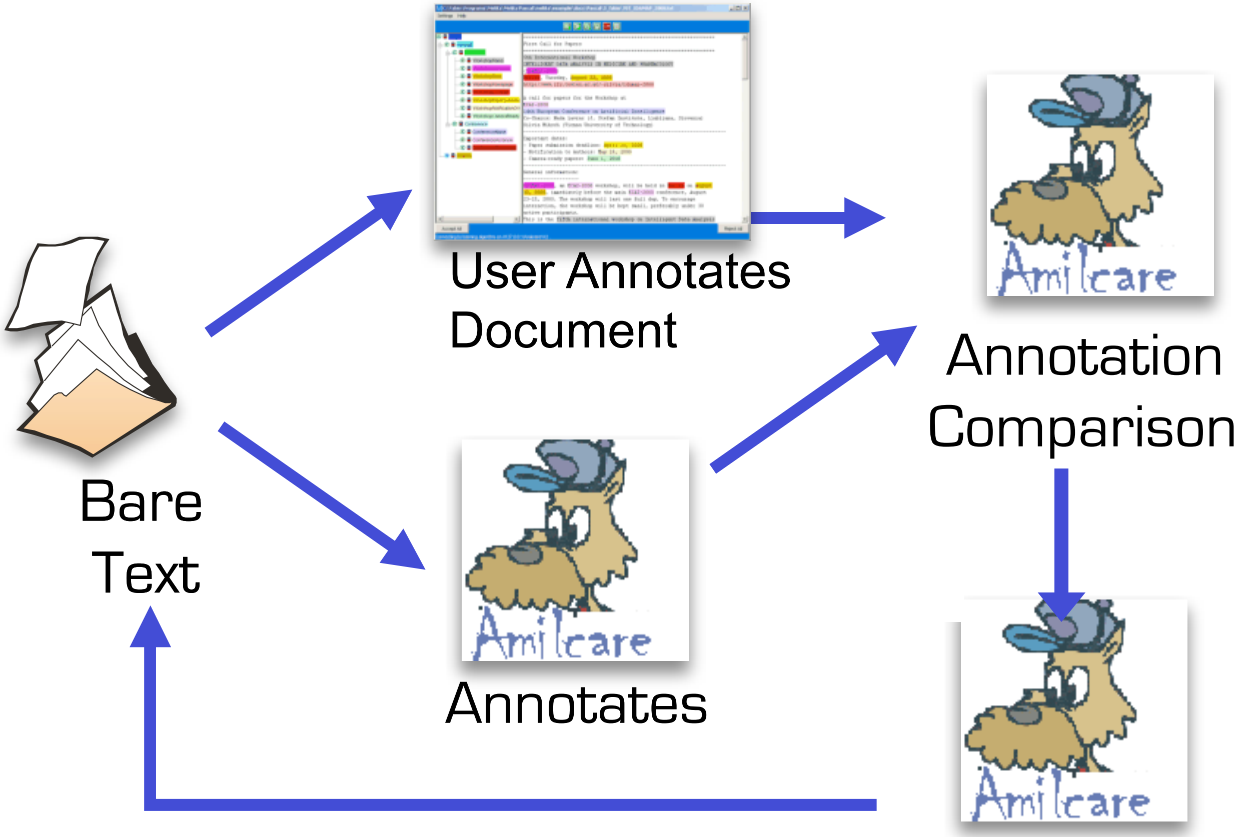


```
Type:     cmu.andrew.academic.sds.seminars
Topic:    SKVORETZ Seminar
Dates:    4-May-95
Time:     4:00 - 5:30
PostedBy: Carole Deamovich on 2-May-95 at 1
Abstract:
Professor John Skvoretz, U. of South Carolin
a seminar entitled "Embedded Commitment," on
4-5:30 in PH 223D.
```

File _name	Suggestion Levels
C:\Program Files\Melita\example\seminar\doc12.txt	88%
C:\Program Files\Melita\example\seminar\doc11.txt	88%
C:\Program Files\Melita\example\seminar\doc13.txt	80%
C:\Program Files\Melita\example\seminar\doc17.txt	79%
C:\Program Files\Melita\example\seminar\doc16.txt	79%
C:\Program Files\Melita\example\seminar\doc14.txt	78%
C:\Program Files\Melita\example\seminar\doc5.txt	76%
C:\Program Files\Melita\example\seminar\doc7.txt	72%
C:\Program Files\Melita\example\seminar\doc2.txt	72%
C:\Program Files\Melita\example\seminar\doc6.txt	58%
C:\Program Files\Melita\example\seminar\doc4.txt	
C:\Program Files\Melita\example\seminar\doc3.txt	
C:\Program Files\Melita\example\seminar\doc1.txt	
C:\Program Files\Melita\example\seminar\doc8.txt	
C:\Program Files\Melita\example\seminar\doc9.txt	
C:\Program Files\Melita\example\seminar\doc10.txt	

Using IE to Support Manual Annotation

Using IE to support annotation: step 1



User Annotates Document

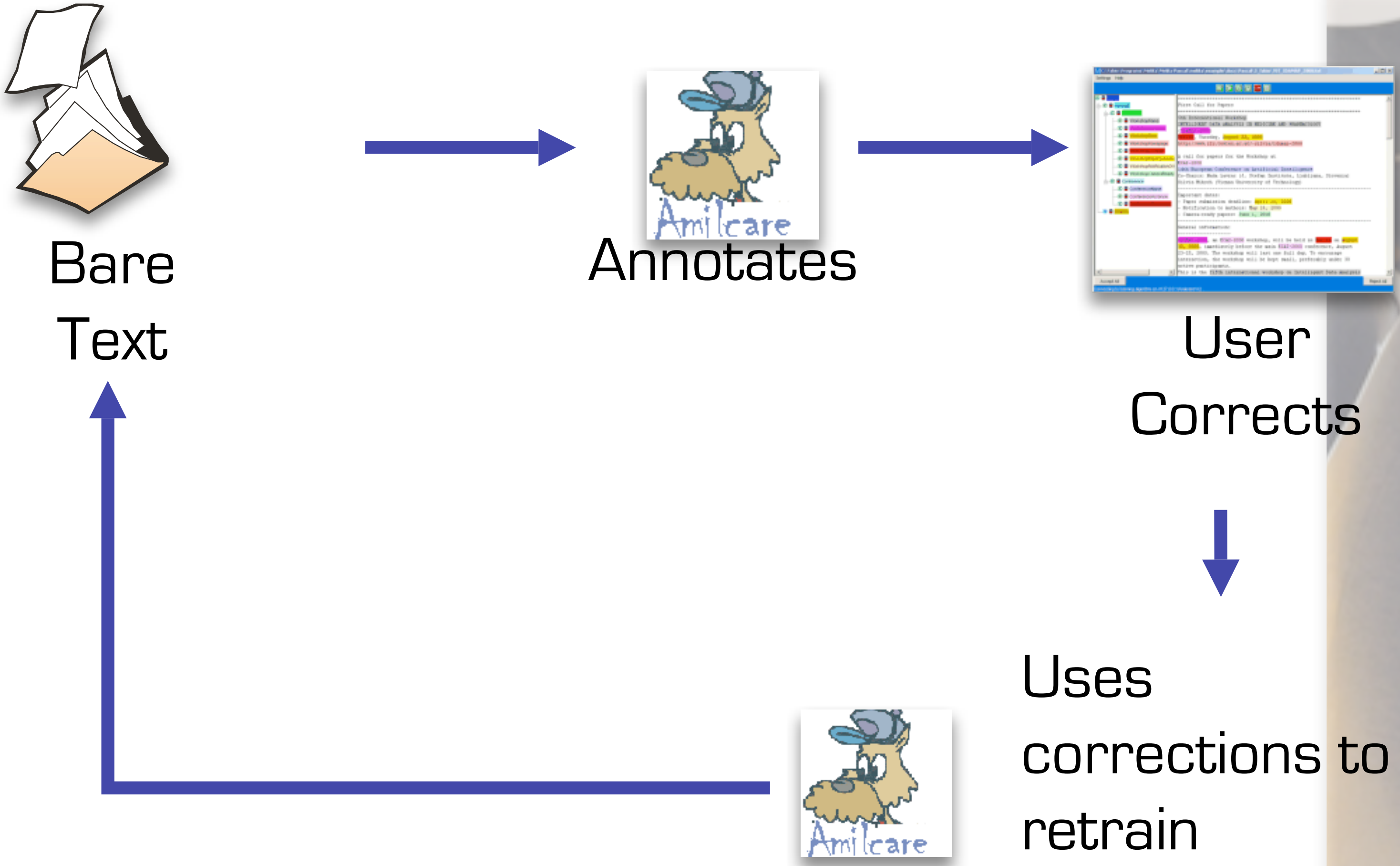
Bare Text

Annotates

Annotation Comparison

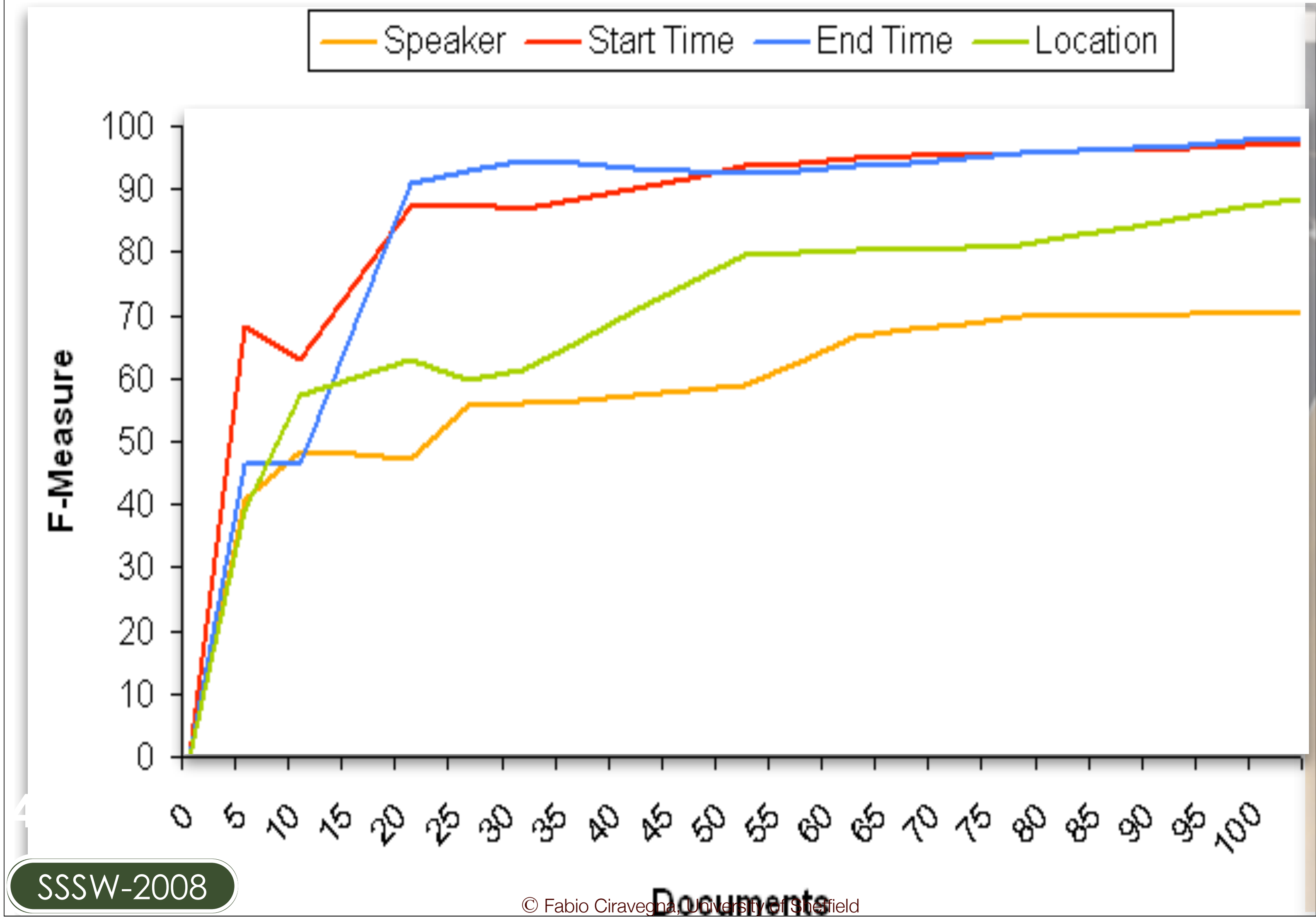
Retrain using errors, missing tags and mistakes

Using IE to support annotation: step 2



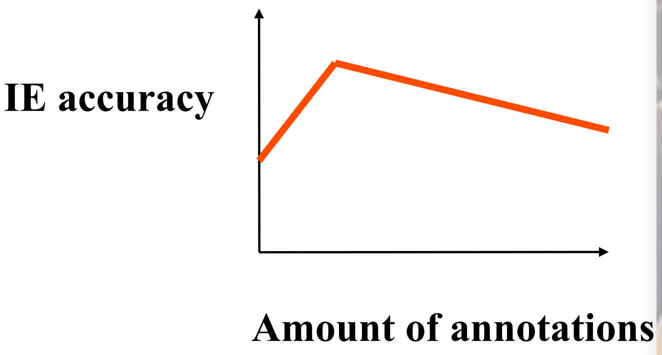
Fabio Ciravegna, Alexiei Dingli, Daniela Petrelli and Yorick Wilks: *User-System Cooperation in Document Annotation based on Information Extraction*, EKAW 2002

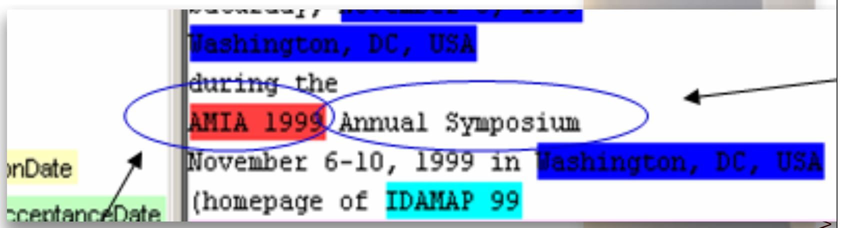
Learning curve



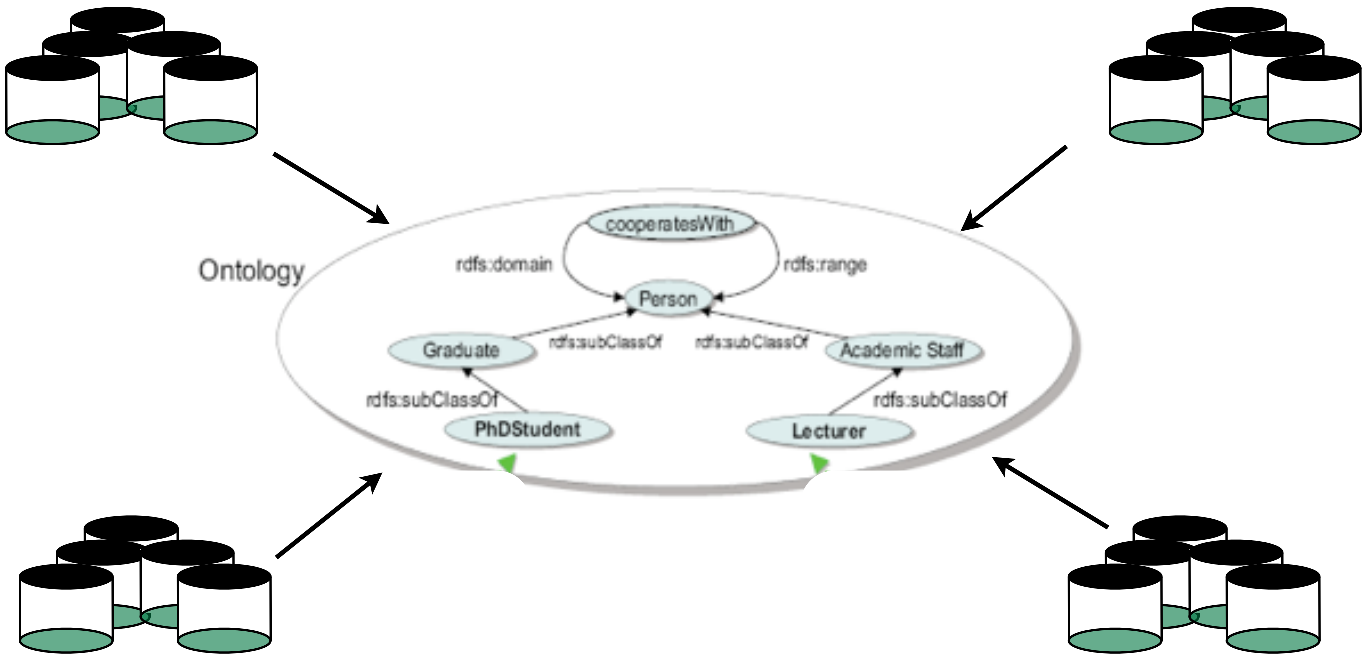
Impact on Annotation

- University of Karlsruhe experiments
  - -80% annotation time
  - +100 interannotator agreement
    - Is this positive?
- Outstanding issue:
  - Impact on annotators of suggestions topping 85% accuracy?
  - Annotation needs to be precise and consistent
    - Otherwise the IE system is confused
    - Can only annotate document content
      - With connections to the rest of the knowledge via information integration





Washington, DC, USA
during the
AMIA 1999 Annual Symposium
November 6-10, 1999 in Washington, DC, USA
(homepage of IDAMAP 99



Information Integration

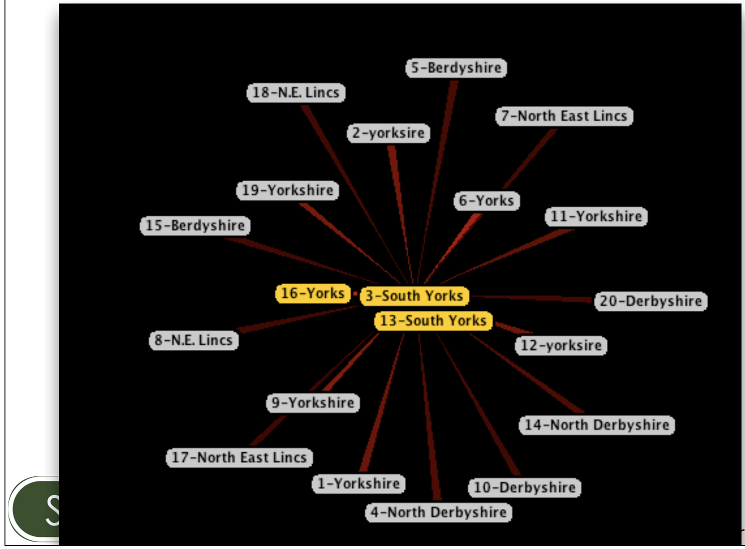
© Fabio Ciravegna, University of Sheffield

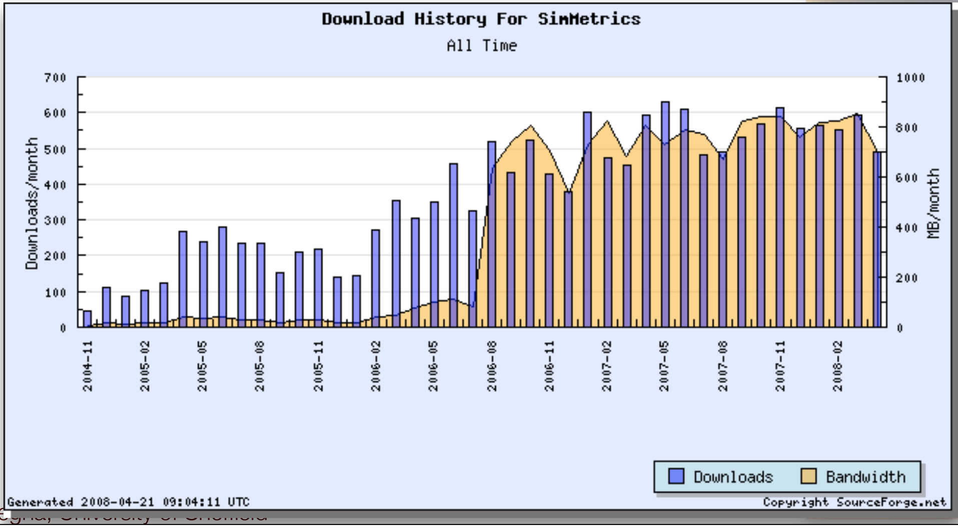
Information Integration

- 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

SimMetrics

- Library of distance metrics released as open source
  - http://sourceforge.net/projects/simmetrics/
  - >15,000 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 Sheffield





Sources

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 database
Guildhall Library

Harben's Dictionary of London

John Strype's "Survey…"

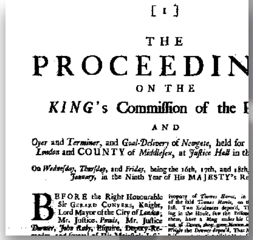
http://www.motco.com

Metropolitan London in the 1690s
IHR

Selected Criminal Records
PRO

House of Lords Journals
BOPCRIS

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



[1]

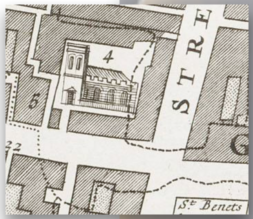
THE

PROCEEDIN

ON THE

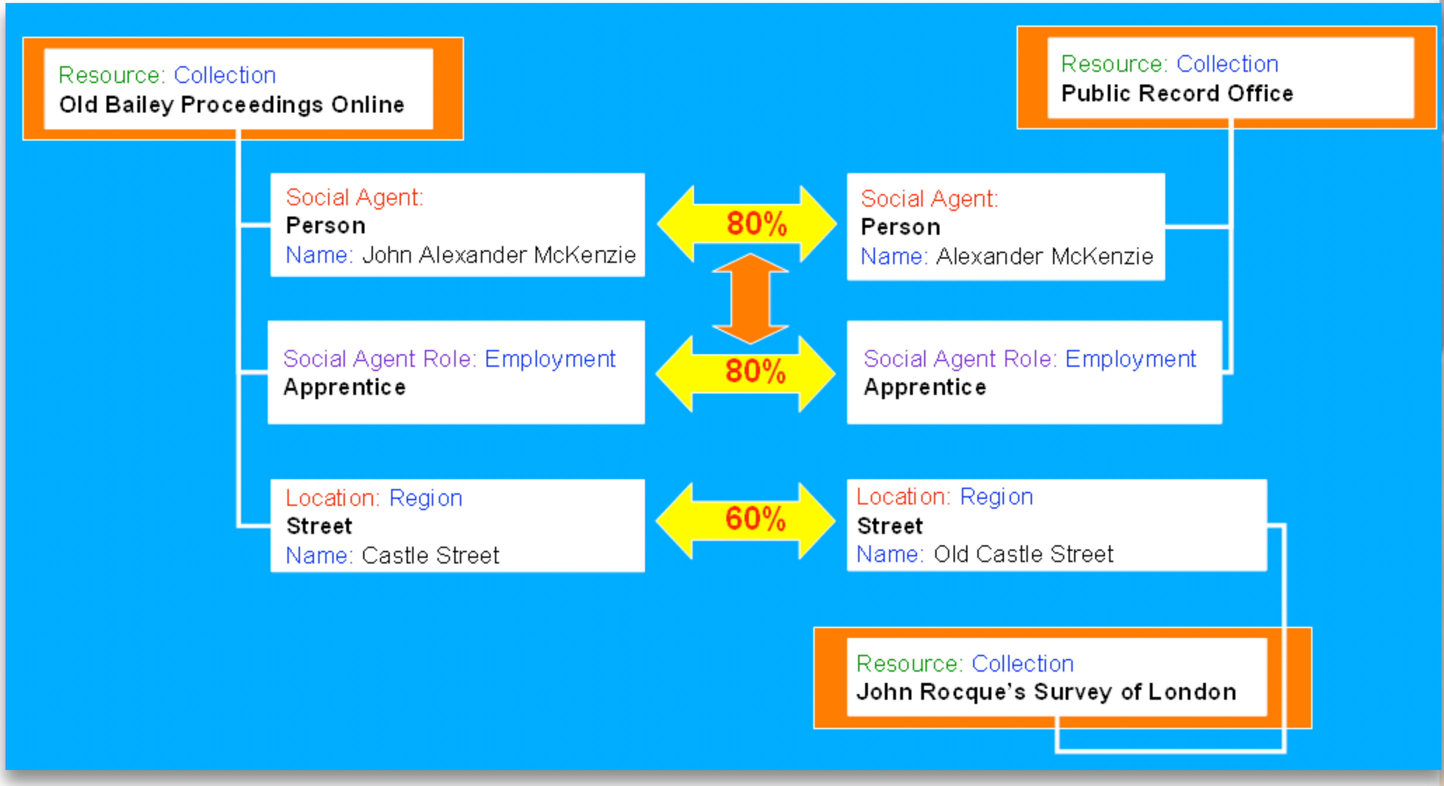
KING's Commiſſion of the

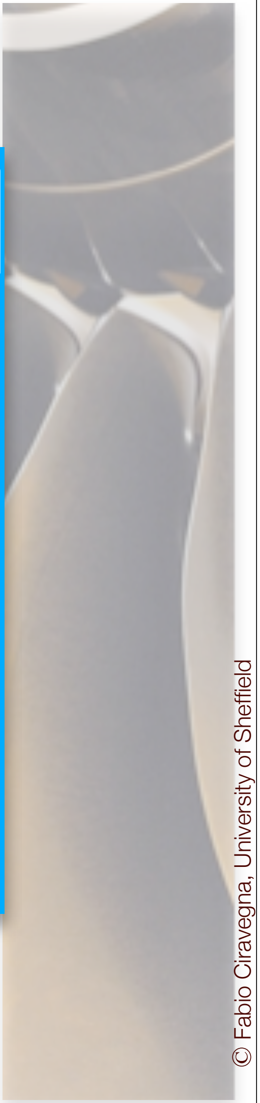
AND



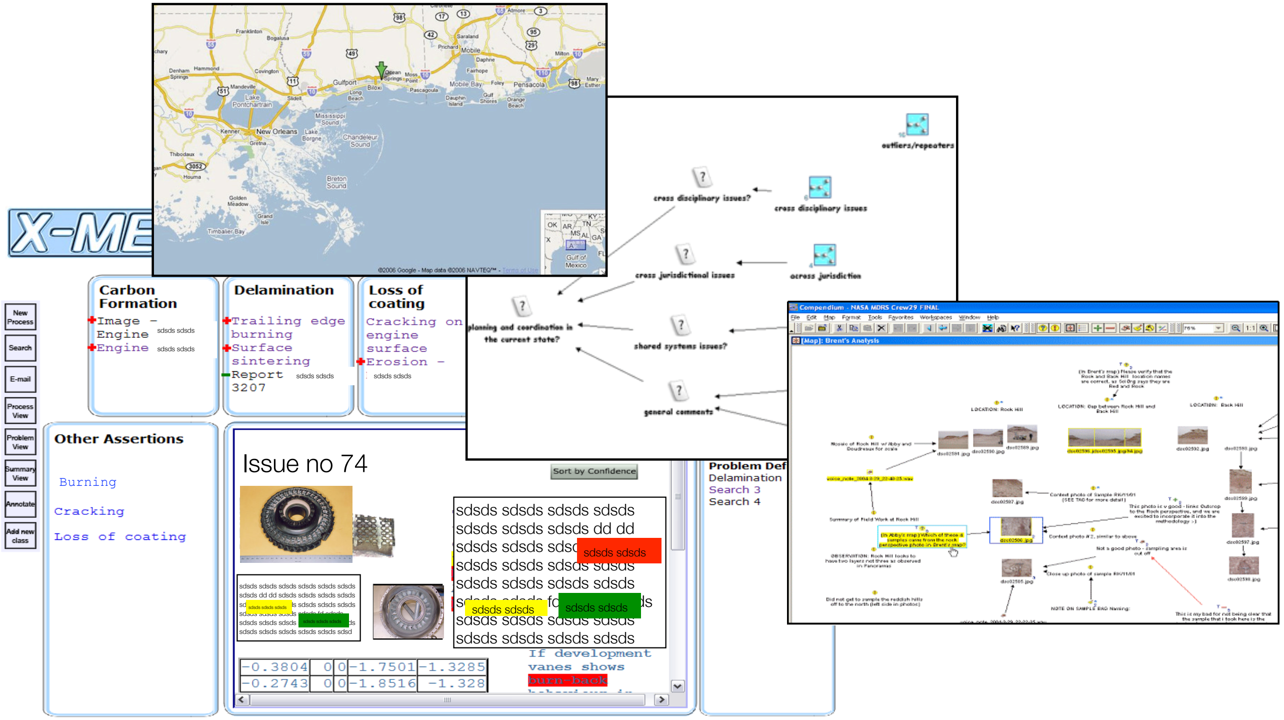
Information Integration

Armadillo: **Historical Data Mining**





Arts & Humanities Research Council

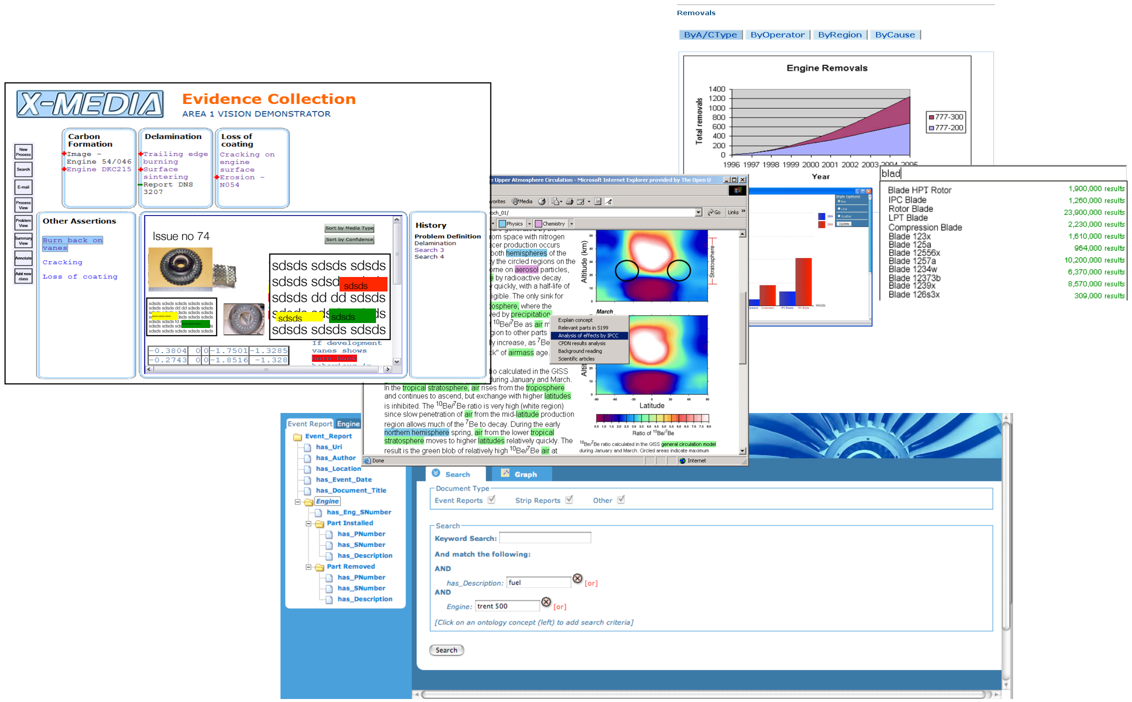


Knowledge Sharing and Reuse

- issues in knowledge sharing
- approaches and novel methods to searching, sharing and reuse knowledge

Knowledge Sharing and Reuse

- In KM mainly means
  - Retrieving information and knowledge
    - At the right time
    - In the right form
        - E.g. independently from where it is stored
        - Or even the form in which it is stored
      - Suitable to the specific users
        - e.g. patients should net receive information using technical terms
      - Suitable to specific interests
        - I am working on social aspects of SW, not interested in engineering aspect of SW
    - In an efficient and effective way
      - Coping with large scale
  - Supporting processes



Document Type
Event Reports ☑ Strip Reports ☑ Other ☑

Search
Keyword Search:
And match the following:
AND
has_Description: fuel [or]
AND
Engine: trent 500 [or]
[Click on an ontology concept (left) to add search criteria]

Search

blad

Blade HPT Rotor	1,900,000 results
IPC Blade	1,260,000 results
Rotor Blade	23,900,000 results
LPT Blade	2,230,000 results
Compression Blade	1,610,000 results
Blade 123x	964,000 results
Blade 125a	10,200,000 results
Blade 12556x	6,370,000 results
Blade 1257a	8,570,000 results
Blade 1234w	309,000 results
Blade 12373b	
Blade 1239x	
Blade 126s3x	

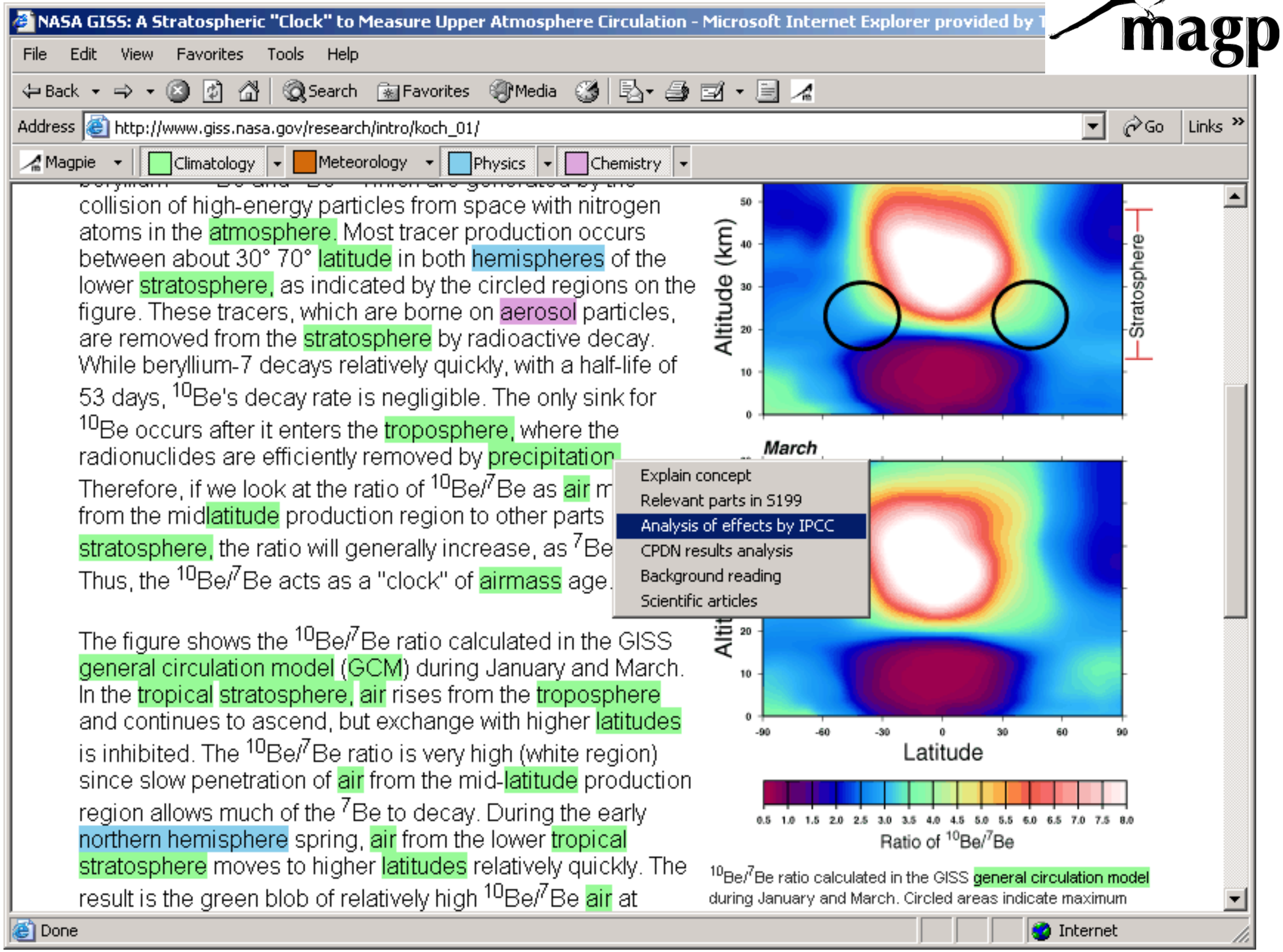
SW for Knowledge Sharing and Reuse

Sharing and Reuse via SW

- Ontology based annotation enables
  - Searching using ontologies
    - Searching metadata rather than text
  - Connection of information across documents, media and archives
    - Retrieving information independently from the store/ media
  - Reasoning on knowledge
    - Making implicit explicit
  - Workflow support
    - Supporting user actions rather than single searches

Document enrichment

- Adding knowledge to documents (ctd.)
  - Document enrichment: helping connecting the document to the rest of the knowledge
    - Associating Services
      - Magpie (Dzbor et al. 2004)
    - Connected to other documents
      - e.g. Automatic generation of hyperlinks
      - COHSE (Goble et al. 2001)


NASA GISS: A Stratospheric "Clock" to Measure Upper Atmosphere Circulation - Microsoft Internet Explorer provided by T
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.giss.nasa.gov/research/intro/koch_01/
Go Links
Magpie Climatology Meteorology Physics Chemistry
Altitude (km)
50 40 30 20 10 0
Stratosphere
March
-90 -60 -30 0 30 60 90
Latitude
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
Ratio of 10Be/7Be
Explain concept
Relevant parts in S199
Analysis of effects by IPCC
CPDN results analysis
Background reading
Scientific articles
Done
Internet

collision of high-energy particles from space with nitrogen atoms in the atmosphere. Most tracer production occurs between about 30° 70° latitude in both hemispheres of the lower stratosphere, as indicated by the circled regions on the figure. These tracers, which are borne on aerosol particles, are removed from the stratosphere by radioactive decay. While beryllium-7 decays relatively quickly, with a half-life of 53 days, ^{10}Be's decay rate is negligible. The only sink for ^{10}Be occurs after it enters the troposphere, where the radionuclides are efficiently removed by precipitation. Therefore, if we look at the ratio of ^{10}Be/^{7}Be as air m from the midlatitude production region to other parts stratosphere, the ratio will generally increase, as ^{7}Be Thus, the ^{10}Be/^{7}Be acts as a "clock" of airmass age.

The figure shows the ^{10}Be/^{7}Be ratio calculated in the GISS general circulation model (GCM) during January and March. In the tropical stratosphere, air rises from the troposphere and continues to ascend, but exchange with higher latitudes is inhibited. The ^{10}Be/^{7}Be ratio is very high (white region) since slow penetration of air from the mid-latitude production region allows much of the ^{7}Be to decay. During the early northern hemisphere spring, air from the lower tropical stratosphere moves to higher latitudes relatively quickly. The result is the green blob of relatively high ^{10}Be/^{7}Be air at

^{10}Be/^{7}Be ratio calculated in the GISS general circulation model during January and March. Circled areas indicate maximum

Searching using Sem Web

- Many types of technologies
  - Search based on structural query languages, such as SPARQL, see, e.g., ARQ, and
  - User-centred search to retrieve ontologies (e.g. Swoogle [Ding et al. 2004] and Watson [d'Aquin et al. 2007])
  - User-centred approaches to retrieve information and knowledge
- We will see the latter

Keywords Based Searching

- KS effectiveness is often affected by two main issues,
  - Ambiguity:
    - Keywords can be polysemous, i.e. they can have multiple meanings.
      - Search returns spurious documents (low precision)
  - Synonymity:
    - an object can be identified by multiple equivalent terms
      - Search does not return documents containing other synonyms (low recall)

Ontology-based Search

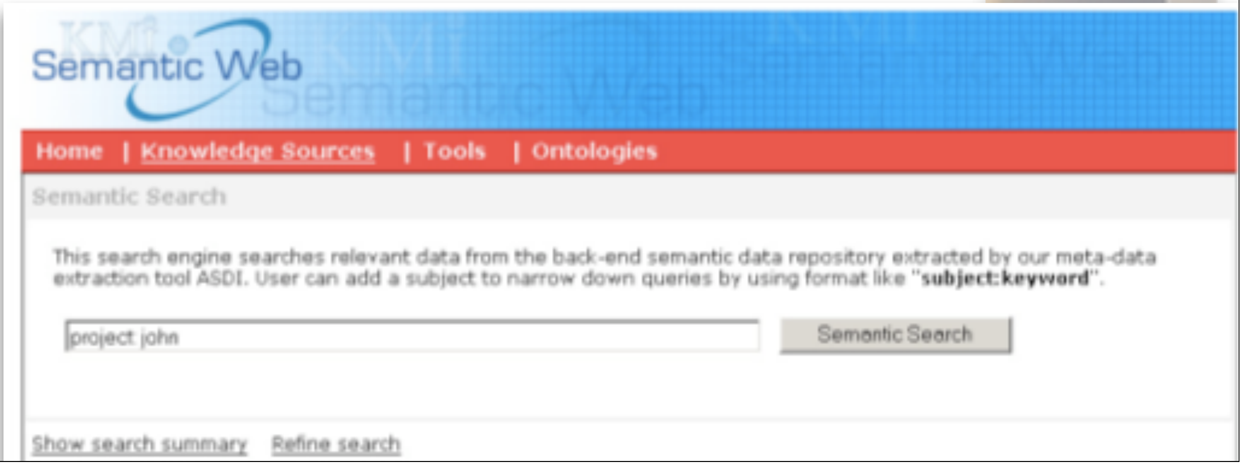
- Searching metadata rather than texts or images
  - Ontology enables reasoning
    - More flexible than searching using traditional methods
- Searching to...
  - Retrieve documents (images/texts/videos/data)
    - As replacement of traditional document management systems
  - Retrieve information/knowledge
    - Querying the knowledge (e.g. the triple store)

User Centred Approaches

- By merging the definitions in [Uren et al. 2008], [Kaufmann et al. 2007b] and [Baghdev et al. 2008]:
  - Keyword-based approaches considering a natural language query as a bag of words
    - [Kaufmann et al. 2007a] [Lei et al., 2006])
  - Natural language approaches: modelling the linguistics of the query
    - [Lopez et al. 2005],[Bernstein et al. 2005b], [Kaufmann et al. 2006]
  - Graph-based approaches
    - [Bernstein et al. 2005a], SEWASIE, Falcon-S.
  - Form-based approaches (e.g. Corese)
  - Hybrid approaches
    - K-Search [Baghdev et al. 2008])

Semantic Search Approaches (1)

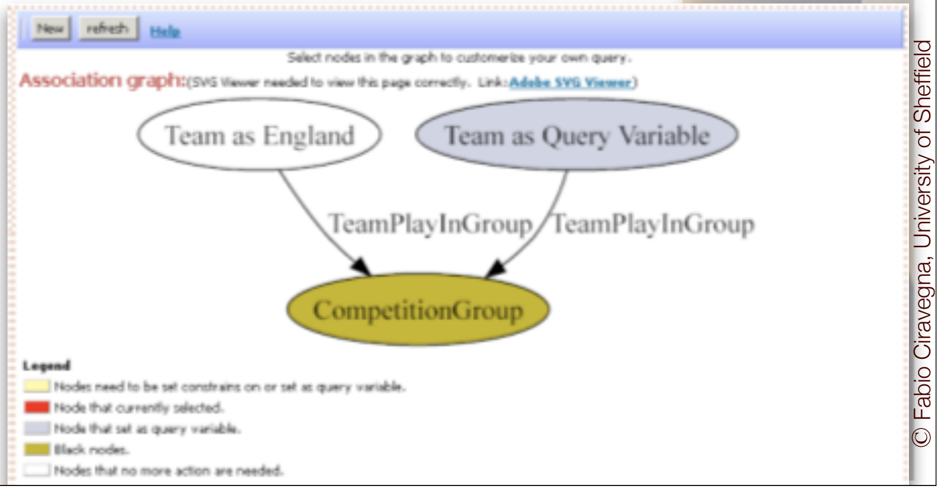
- Keyword-based approaches
  - Query via keywords
  - All the keywords are mapped to Semantic Concepts
  - Requirements: feedback on generated query
  - Issues:
    - User lost for words
      - What is covered by the ontology?
- E.g. SemSearch



This search engine searches relevant data from the back-end semantic data repository extracted by our meta-data extraction tool ASDI. User can add a subject to narrow down queries by using format like "**subject:keyword**".

Semantic Search Approaches (2)

- View-based approaches
  - Based on querying by building visual graphs
  - Advantages:
    - What covered by ontology is always clear
  - Issues
    - Can be fairly rigid and constraining
    - Kaufmann et al 2007 report a very high time required for querying
- E.g. Falcon



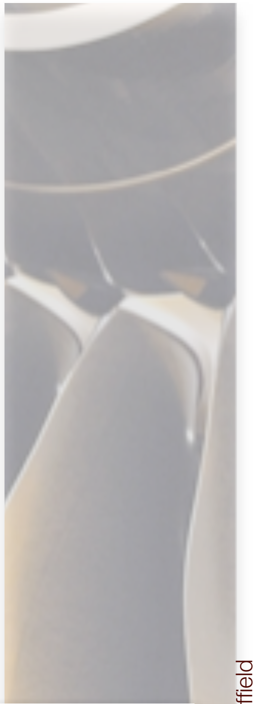
Select nodes in the graph to customerize your own query.

Association graph:(SVG Viewer needed to view this page correctly. Link:Adobe SVG Viewer)

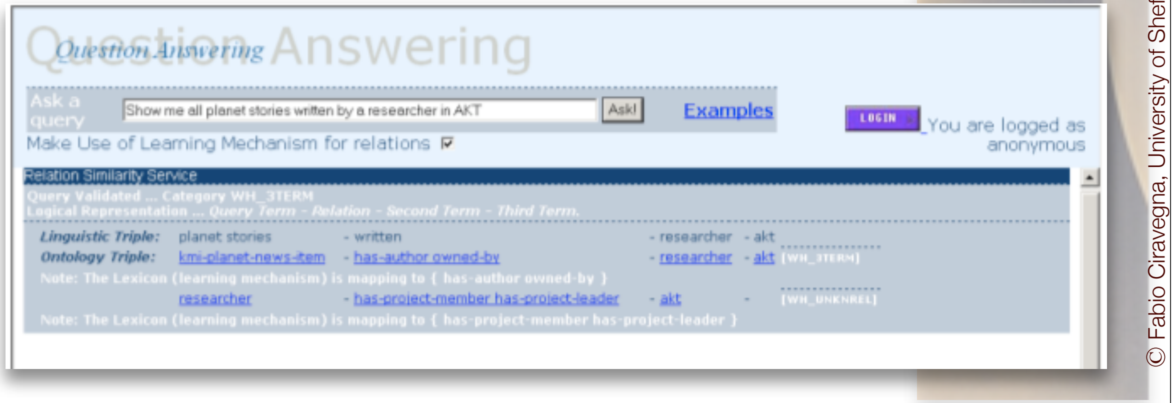
Legend

- 🟨 Nodes need to be set constrains on or set as query variable.
- 🟥 Node that currently selected.
- ▨ Node that set as query variable.
- 🟨 Black nodes.
- □ Nodes that no more action are needed.

Semantic Search Approaches (3)



- A natural language approach
  - Interprets full fledged NL questions
  - Requirements:
    - Feedback on generated query
  - Issues:
    - User lost for words
      - What is covered by the ontology?
    - NL can be tricky (linguistic coverage)
- E.g. Aqua



Show me all planet stories written by a researcher in AKT

Make Use of Learning Mechanism for relations ☑

Relation Similarity Service

Query Validated ... Category WH_3TERM
Logical Representation ... *Query Term - Relation - Second Term - Third Term.*

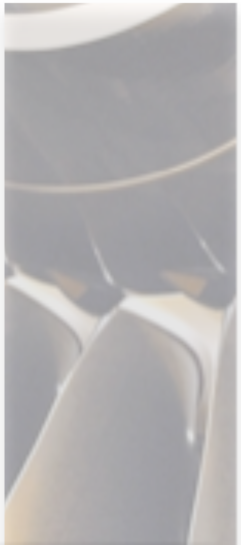
Linguistic Triple:	planet stories	- written	- researcher	- akt	
Ontology Triple:	kmi-planet-news-item	- has-author owned-by	- researcher	- akt	[WH_3TERM]

Note: The Lexicon (learning mechanism) is mapping to { has-author owned-by }

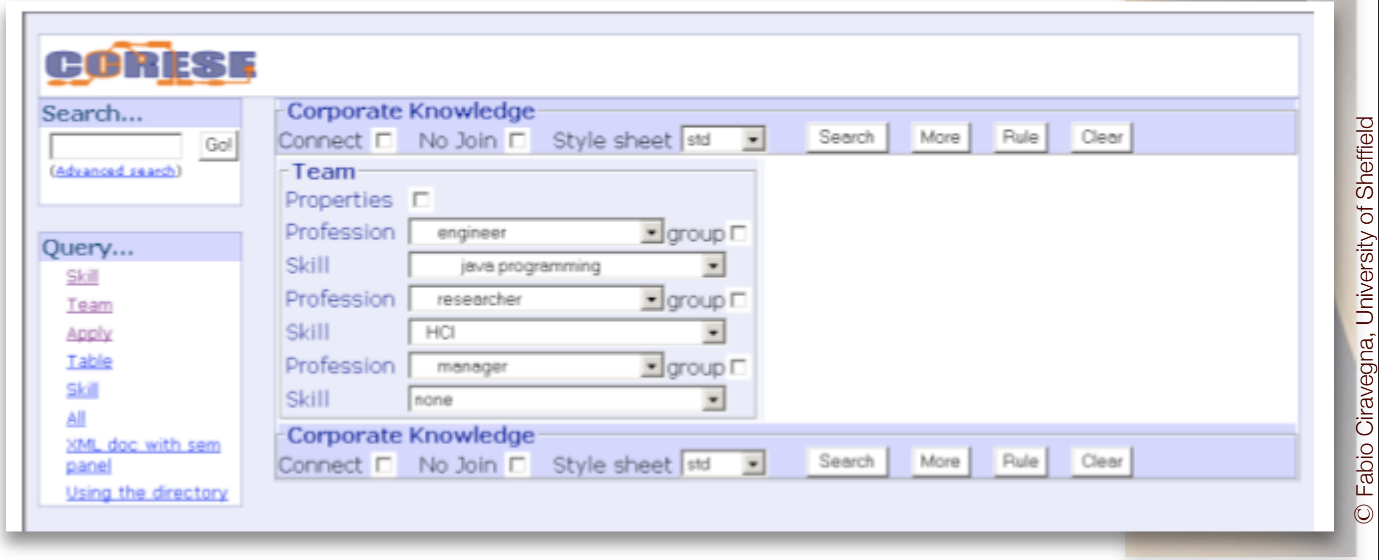
	researcher	- has-project-member has-project-leader	- akt	-	[WH_UNKNREL]

Note: The Lexicon (learning mechanism) is mapping to { has-project-member has-project-leader }

Semantic Search Approaches (4)



- Form-based approaches
  - The ontology is turned into a form and queries are expressed by filling conditions into the form
    - Advantages:
      - What covered by ontology is always clear
    - Issues
      - Can be fairly rigid and constraining



Corporate Knowledge

Connect □ No Join □ Style sheet std

Team

Properties □

Profession engineer group □

Skill java programming

Profession researcher group □

Skill HCI

Profession manager group □

Skill none

Corporate Knowledge

Connect □ No Join □ Style sheet std

Ontology-based Querying: Issues

- Metadata may cover only partially the user information needs
  - Limitations in the ontology wrt user needs
    - Often the use people will do of information is impossible to foresee
  - Limitations in the annotation capabilities
    - Sometimes Information is impossible to retrieve reliably using automatic methods
  - Metadata unavailable for a specific document

An Experiment on Jet Engine Event Reports

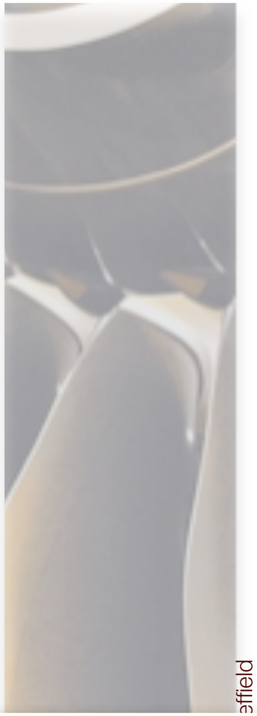
- 21 topics of search, e.g.
  - "How many events were caused during maintenance in 2003?"
  - "What events were caused during maintenance in 2003 due to control units?"
  - 'Find al l the events associated with damage to acous- tic liners fol lowing bird strike"

- How many topics can we model with Information Extraction?
  - 21 topics/ 14 topics partially or not covered by IE-based annotations
    - given size of corpus there is no way that manual annotations are added

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

Hybrid Search



- [Bhagdev et al 2008] propose a model of searching combining
  - the flexibility of keyword-based retrieval
  - querying and reasoning capabilities of semantic search
- HS is formally defined as:
  - the application of semantic (metadata-based) search for the parts of the user queries
    - where metadata is available
  - the application of keyword-based covered by metadata.

- But also it must leave freedom to users to chose among the two paradigms!
  - As we will see users make a creative use of it

Queries in Hybrid Search

differently from other approaches (e.g. [9]), in HS conditions on metadata and keywords coexist.

- Any boolean combination of three type conditions
  - pure semantic:
    - via unique identification of objects/relations
      - e.g. via URIs or unique identifiers
  - keyword-based
    - matching on the whole document
  - keyword-in-context
    - e.g. it enables searching for the string "fuel" but only in the context of all the text portions annotated with the concept affected-engine-part [14]

Example of Hybrid Query

∀x,y,z /

(discoloration y) & (located-on y x) & (component x)

Querying Metadata

& (provenance-text-contains x “blade”)

Keyword in Context Query

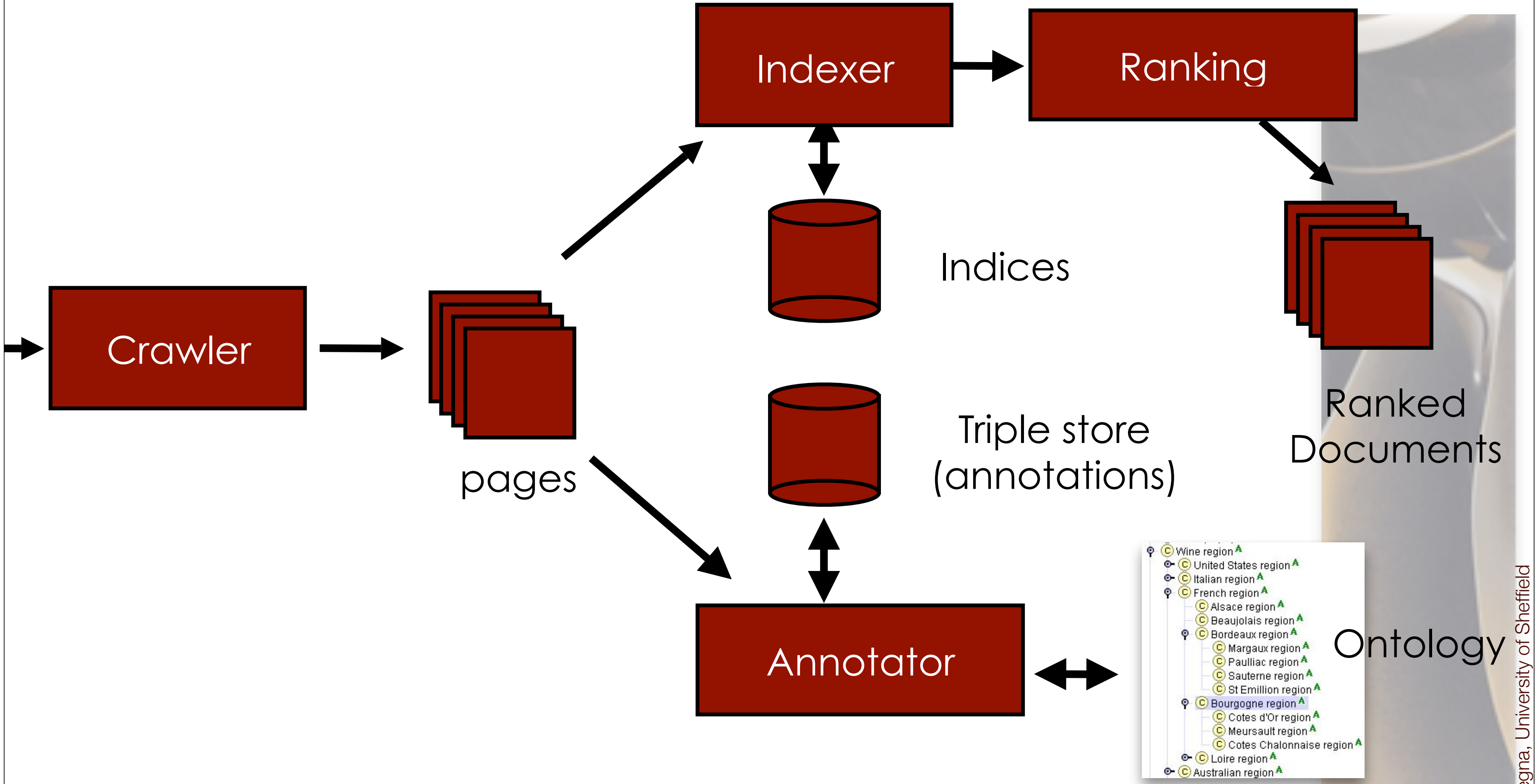
& (contains z “trailing edge”) & (document z) & (provenance x z)

Keyword-based Query

Implementing HS: Indexing

- Documents are indexed using a standard keyword–based engine such as SolR
- Facts (e.g. extracted by an IE system) are stored in a Knowledge Base
  - e.g. a triple store like Sesame2 in the form of RDF triples.
- Provenance of facts recorded
  - E.g. As triples connecting
    - the facts' URIs and those of the document of origin
    - the facts' URIs and the original strings used in the documents

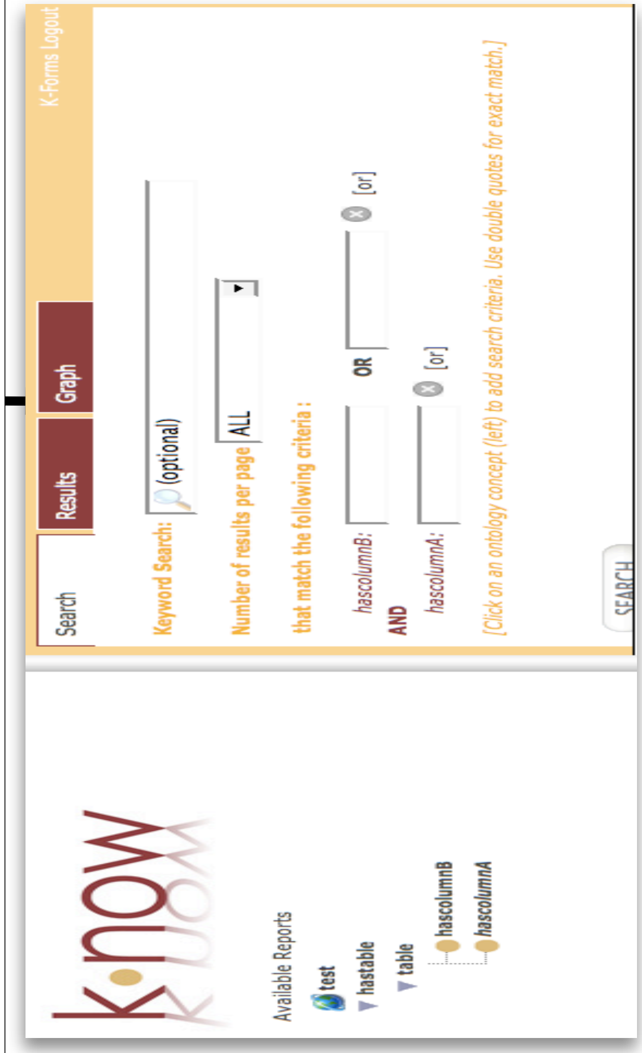
K-Search: indexing



Implementing HS: Retrieval

- Query is parsed and the different components (keywords, keywords-in-context and metadata) identified
  - keyword matches ➜ traditional information retrieval system
  - metadata searches
    - Translated into a query language like SPARQL
    - Sent to a triple store
  - keywords-in-context queries
    - matched with provenance of annotations in documents
      - E.g. Using SPARQL and a triple store
- Finally, results are merged, ranked and displayed

K-Search: retrieval



K-Forms Logout

Keyword Search: (optional)

Number of results per page ALL

that match the following criteria :

hascolumnB: OR [or]

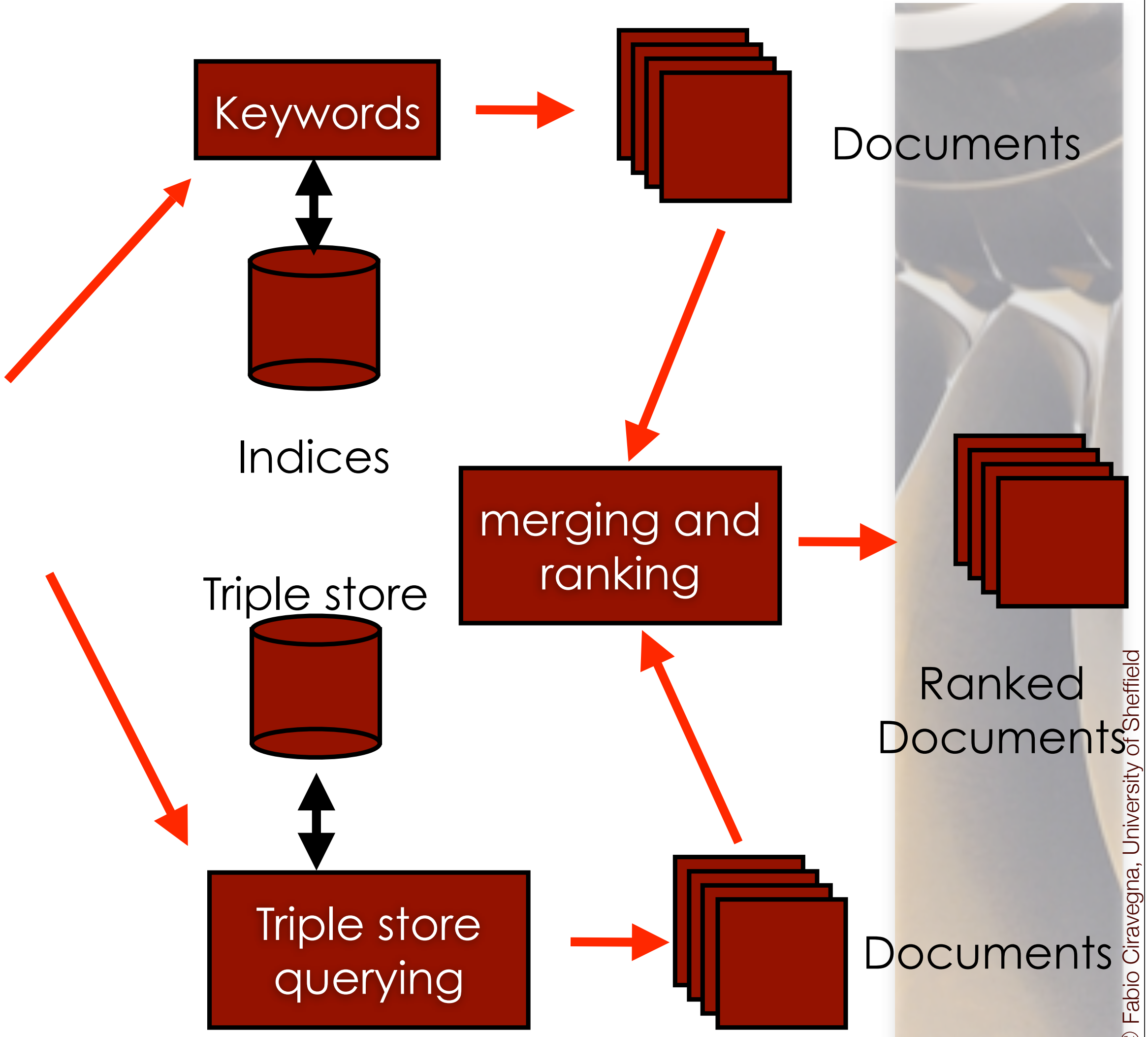
AND

hascolumnA: [or]

[Click on an ontology concept (left) to add search criteria. Use double quotes for exact match.]

Available Reports

- test
  - hastable
    - table
      - hascolumnB
      - hascolumnA



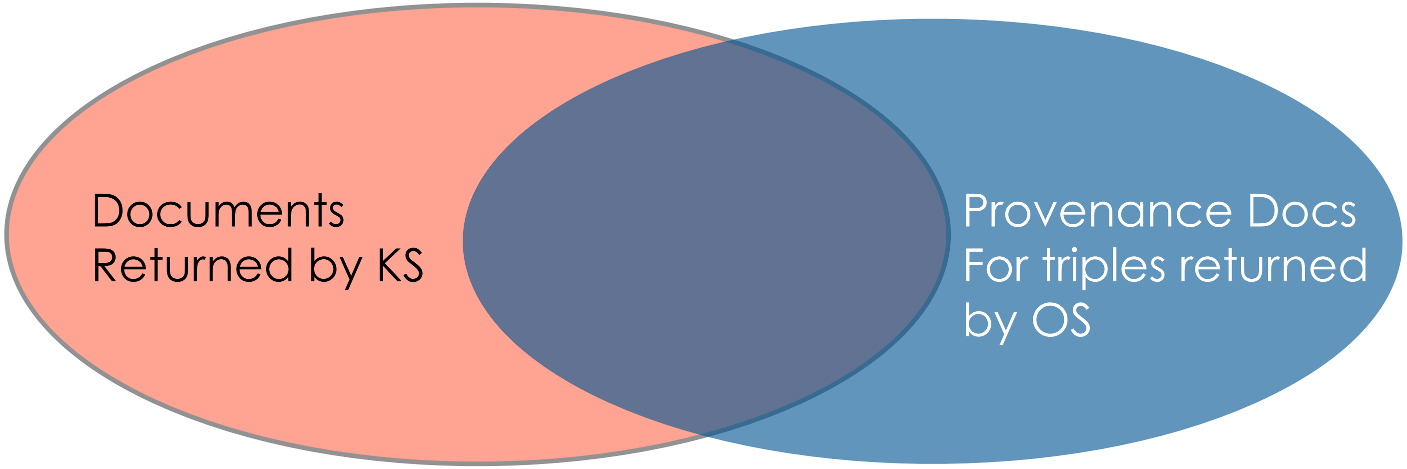
Result Merging

- Merging keyword and semantic results is not straightforward
  - Keyword matching returns an ordered set of URIs of documents
  - a semantic search returns an unordered set of assertions < subj, rel, obj>
- Merging is a different task if:
  - Document Searching
    - Returns documents
  - Knowledge Searching
    - Returns triples

Merging results

- Provenance of triples returns document ids for triples (URIs)
  - Document Searching:
    - Provenance URI set is intersected with URIs of documents returned by keywords
    - `HybridSearchUriSet= KSDocUriSet ∩ OSDocUriSet`

I won’t mention ranking here

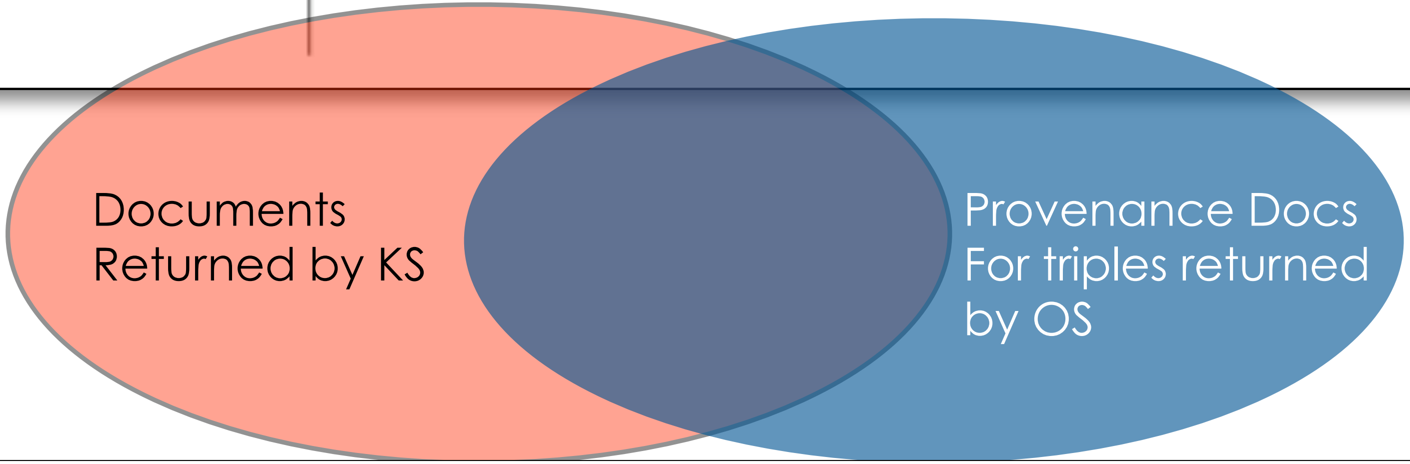


Merging results

- Provenance of triples returns document ids for triples (URIs)
  - Knowledge Searching
    - Triples returned by semantic search are filtered to remove those whose provenance does not point to any of the documents returned by the keywords

I won't mention ranking here

```
HSTripleSet = All triples ∈ OSTripleSet
              Where Provenance(triple^i) ∈ KSDocUriSet
```



Ranking for Document retrieval

- Effective ranking is extremely important for a positive user experience
- Different ranking methods are possible
  - Document based
    - ability to match the keyword-based query
    - the keywords used in anchor links
    - the document popularity (given by link-based weights)
  - Knowledge Based
    - Presence and quality of metadata

Expected effect of HS: Document Searching

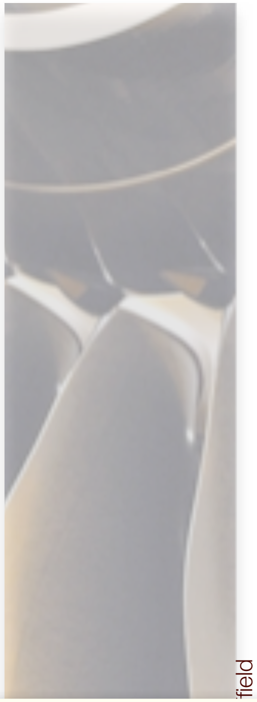
- With respect to OS
  - Recall expected to increase
    - Use of keywords where metadata is missing enables to answer otherwise impossible queries
  - Precision may suffer because of polysemy
- With respect to KS
  - Precision and recall expected to increase
    - Ambiguity and synonymity are dealt with by semantic search when available
      - Higher recall and precision
    - As keywords are combined with metadata in the same query, the context given by the available metadata helps in disambiguating keywords as well
      - higher precision

Expected effect of HS : Knowledge Searching

- With respect to OS
  - Precision increased
    - Use of keywords where metadata is missing enables more precise queries
      - although less precise than the ideal ones

OR

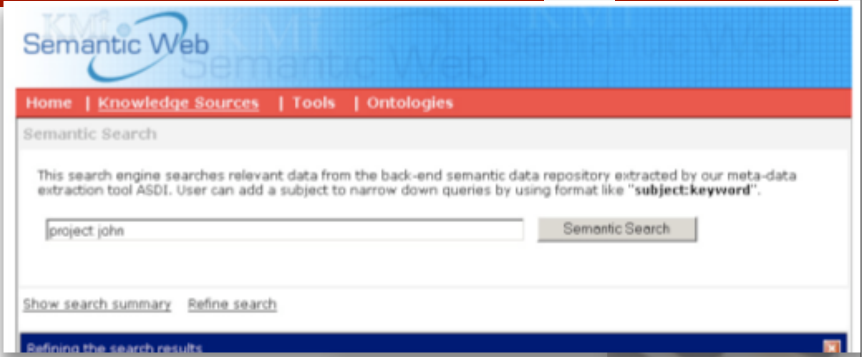
  - Recall increased
    - Use of keywords where metadata is missing enables to answer otherwise impossible queries
  - Precision may suffer because of polysemy
- With respect to KS
  - KS does not cover Knowledge Searching



field

Next slide:
We have implemented a version to confirm our expectation

Implementing HS: What Search Strategy?

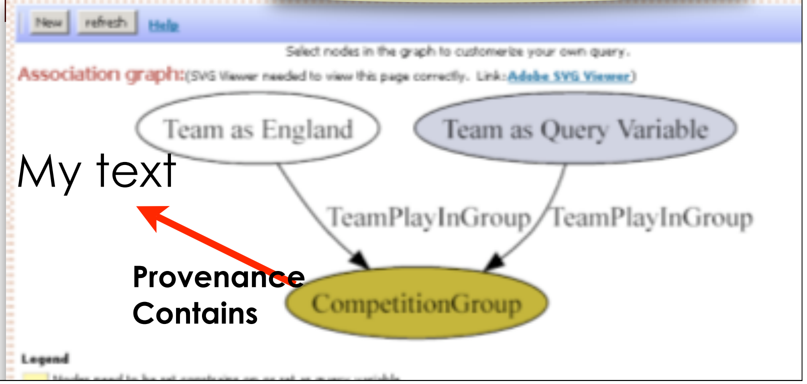


This search engine searches relevant data from the back-end semantic data repository extracted by our meta-data extraction tool ASDI. User can add a subject to narrow down queries by using format like "**subject:keyword**".

- Keyword-based approaches
  - Require translating all the keywords in order to perform the query
    - E.g. SemSearch
    - HS implemented by replacing keywords in the query with c the ontology when possible while leaving the rest for pure k based searching
    - Keywords in context rather difficult

Go through this and next slide very quickly !!

- View-based approaches
  - Based on querying by building visual graphs
    - E.g. Falcon
    - HS support by adding two arc types
      - document-contains
      - Object description contains

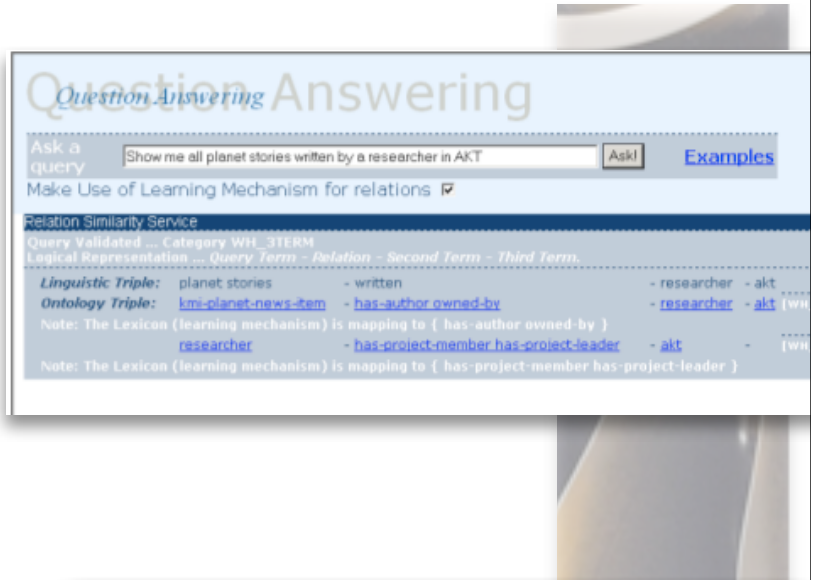


Select nodes in the graph to customize your own query.

Association graph: (SVG Viewer needed to view this page correctly. Link: Adobe SVG Viewer)

Search Strategy (ctd)

- A natural language approach
  - E.g. Aqua
  - HS suported by recognising expressions like
    - “and the document contains…”
    - And its description contains



Question Answering

Ask a query: Show me all planet stories written by a researcher in AKT

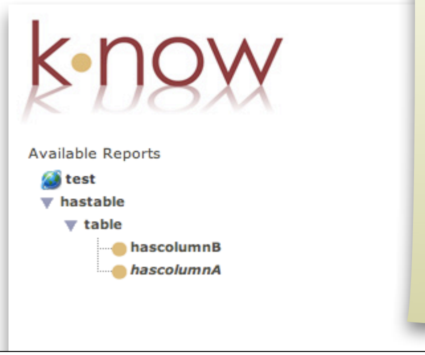
Make Use of Learning Mechanism for relations ☑

Relation Similarity Service

Query Validated ... Category WH_3TERM
Logical Representation ... *Query Term - Relation - Second Term - Third Term.*

Linguistic Triple: planet stories - written - researcher - akt
Ontology Triple: kmi-planet-news-item - has-author owned-by - researcher - akt
Note: The Lexicon (learning mechanism) is mapping to { has-author owned-by }
researcher - has-project-member has-project-leader - akt -
Note: The Lexicon (learning mechanism) is mapping to { has-project-member has-project-leader }

- Form-based approaches
  - HS supported by introducing
    - Keyword Search field
    - Enable keyword Matching on fields



Available Reports
- test
- hastable
  - table
    - hascolumnB
    - *hascolumnA*

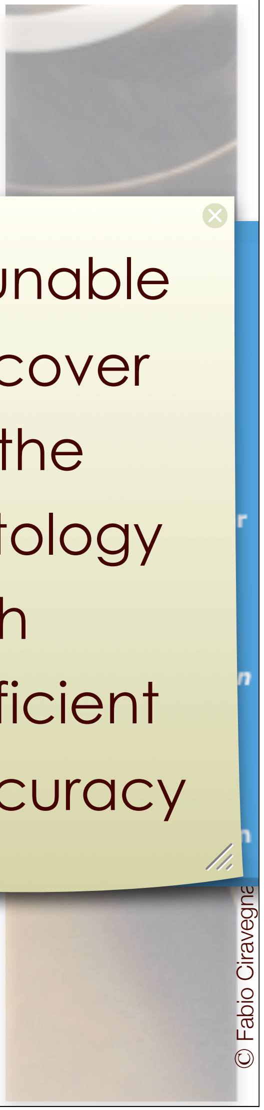
- Form-based implementation of hybrid search initially created for Jet Engine Designers

Putting Everything Together

An experience in the aerospace domain

Annotating Documents

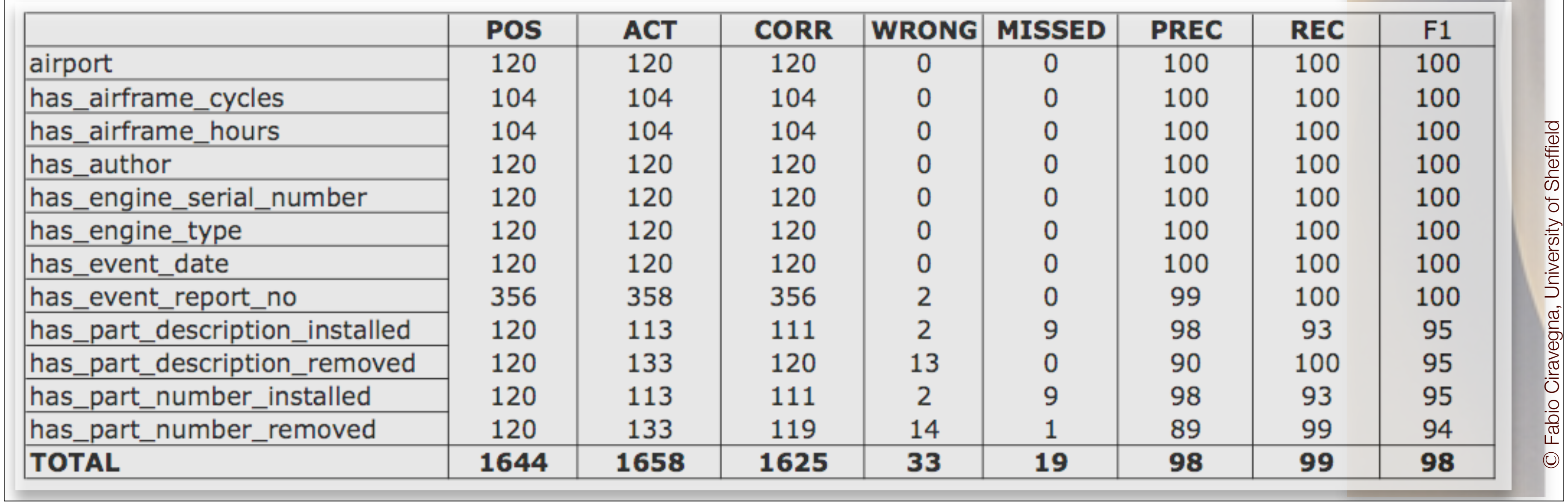
- Automatic extraction of information from event report
  - 18,000 documents analysed
    - Mainly Forms implemented in Word
- Metadata generated according to an ontology developed by Aberdeen U
  - Examples manually annotated by users using AktiveMedia
  - Machine Learning + HLT (T-Rex platform) to train the system to annotate
- Automatic extraction of metadata and indexing of documents



IE unable to cover all the ontology with sufficient accuracy

Applying information extraction

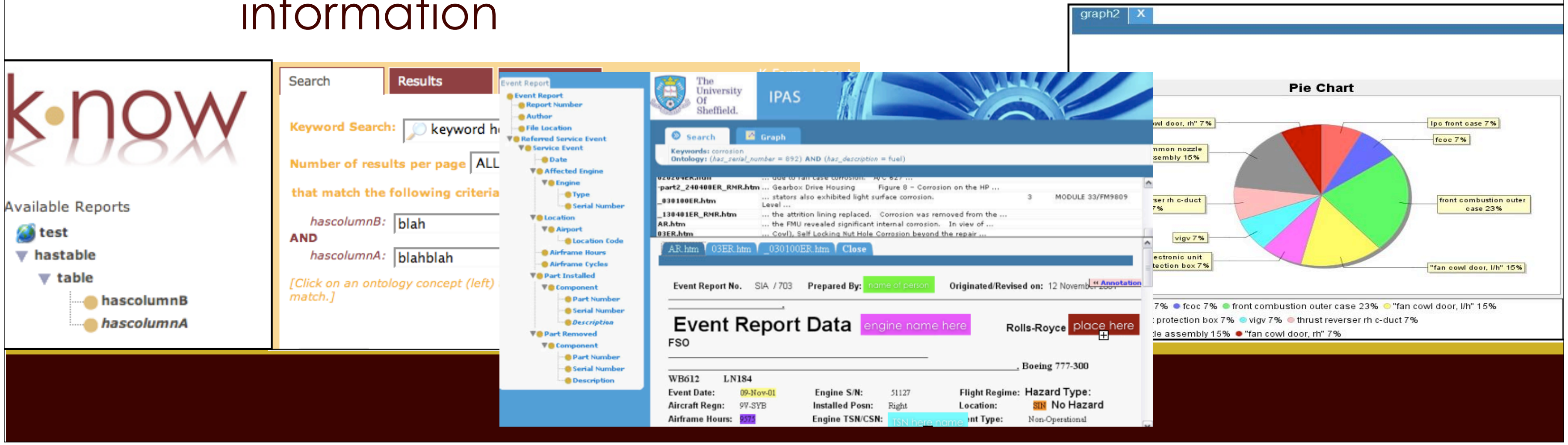
- AktiveMedia to annotate texts
- TRex system (Jiria et al. 2006) to train and extract
    - http://tyne.shef.ac.uk/t-rex/
- IE captures all the information in tables
    - 99% of the information captured (recall=99)
    - 98% of proposed information is correct (precision=98)



	POS	ACT	CORR	WRONG	MISSED	PREC	REC	F1
airport	120	120	120	0	0	100	100	100
has_airframe_cycles	104	104	104	0	0	100	100	100
has_airframe_hours	104	104	104	0	0	100	100	100
has_author	120	120	120	0	0	100	100	100
has_engine_serial_number	120	120	120	0	0	100	100	100
has_engine_type	120	120	120	0	0	100	100	100
has_event_date	120	120	120	0	0	100	100	100
has_event_report_no	356	358	356	2	0	99	100	100
has_part_description_installed	120	113	111	2	9	98	93	95
has_part_description_removed	120	133	120	13	0	90	100	95
has_part_number_installed	120	113	111	2	9	98	93	95
has_part_number_removed	120	133	119	14	1	89	99	94
TOTAL	**1644**	**1658**	**1625**	**33**	**19**	**98**	**99**	**98**

K-Search

- Form-based implementation of hybrid search initially created for Jet Engine Designers
- It enables
  - Document querying
  - Knowledge querying
    - Including quantification of unstructured information



Keyword Search: keyword h

Number of results per page ALL

that match the following criteria

hascolumnB: blah

AND

hascolumnA: blahblah

[Click on an ontology concept (left)
match.]

Keywords: corrosion
Ontology: (has_serial_number = 892) AND (has_description = fuel)

```
-part2_240400ER_RMR.htm ... Gearbox Drive Housing      Figure 8 – Corrosion on the HP ...
_030100ER.htm           ... stators also exhibited light surface corrosion.      3     MODULE 33/FM9809
                        Level ...
_130401ER_RMR.htm       ... the attrition lining replaced.   Corrosion was removed from the ...
AR.htm                  ... the FMU revealed significant internal corrosion.   In view of ...
03ER.htm                ... Cowl), Self Locking Nut Hole Corrosion beyond the repair ...
```

AR.htm | 03ER.htm | _030100ER.htm | Close

Event Report No. SIA / 703 Prepared By: name of person Originated/Revised on: 12 Novemb « Annotation

Event Report Data engine name here Rolls-Royce place here

FSO

, Boeing 777-300

WB612 LN184

Event Date:	09-Nov-01	Engine S/N:	51127	Flight Regime:	Hazard Type:
Aircraft Regn:	9V-SYB	Installed Posn:	Right	Location:	SIN No Hazard
Airframe Hours:	9575	Engine TSN/CSN:	TSN here name	nt Type:	Non-Operational

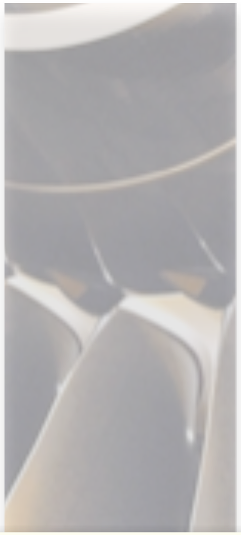
K-Search evaluation

- We have performed 2 types of technology evaluations using K-Search:
  - in vitro:
    - Effectiveness of annotation and query strategy with respect to standard KS and OS
  - in vivo: testing the system with real users
    - 32 users Rolls-Royce engineers
      - Evaluation enables verifying suitability for use in a real environment

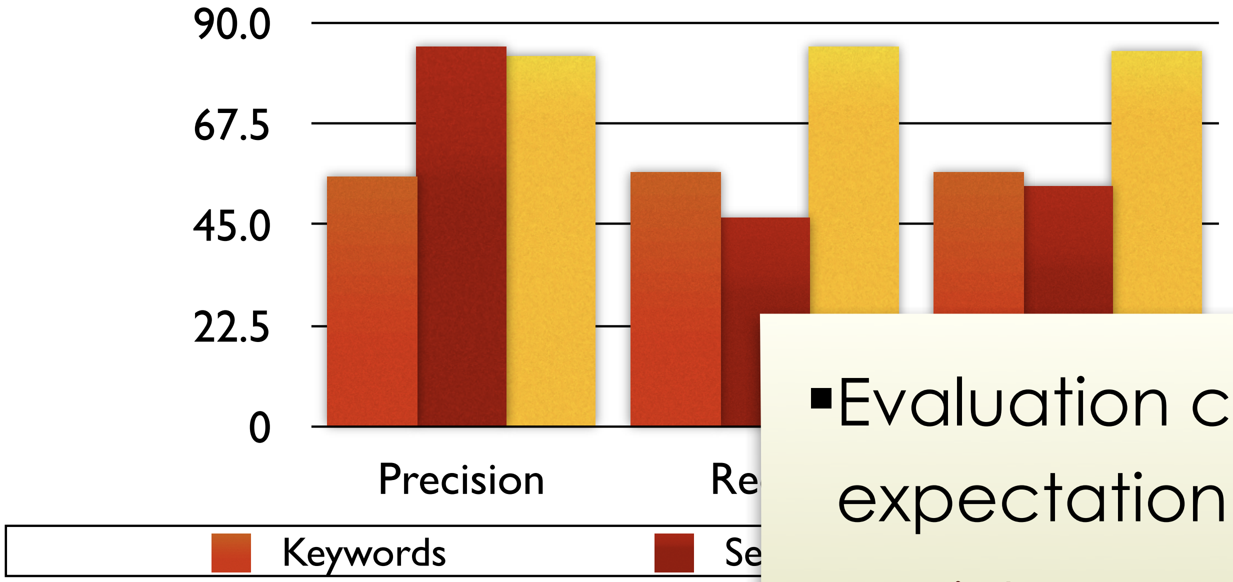
In vitro evaluation

- 21 topics of search, discussed with users, e.g.
  - "How many events were caused during maintenance in 2003?"
  - "What events were caused during maintenance in 2003 due to control units?"
  - 'Find al l the events associated with damage to acous- tic liners fol lowing bird strike"
- Queries:
  - "what events caused during maintenance in 2003 were due to control units?"
- Translated into a set of queries in KS, OS and HS

K-Search on Event Reports



- Accuracy in the first 20 hits on a sample of 400 docs



- Similar results for 50 hits

- Evaluation confirms our expectation:
  - Higher recall wrt OS and KS
  - Higher precision wrt KS
  - Slightly lower precision wrt OS

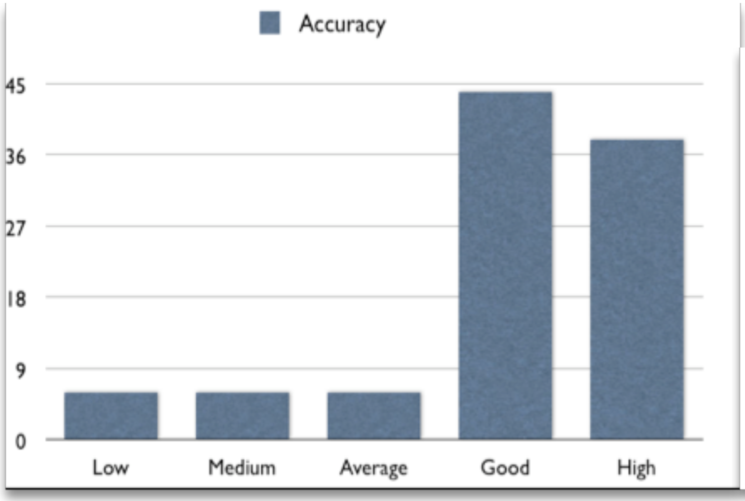
Final User Evaluation

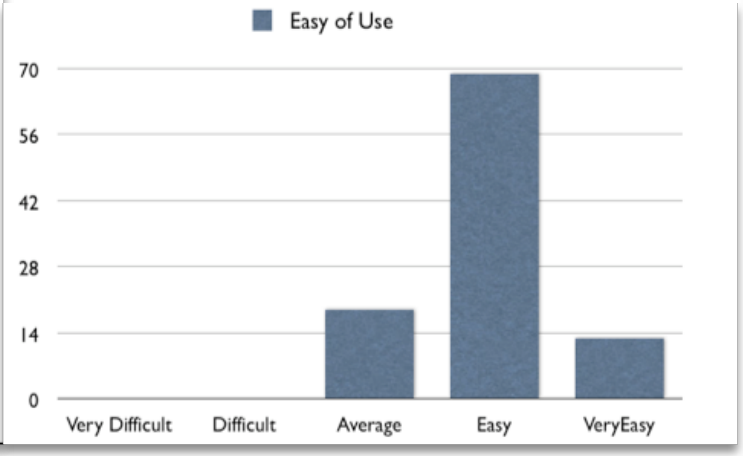
- Goal: verifying suitability for use in a real environment
  - 32 users Rolls-Royce engineers from different parts of the company
  - 90 minutes of test
    - Short introduction
    - 3 monitored tasks
      - One given (including solution)
      - One given (no solution)
      - One free task
  - Availability of system on intranet for the following period
- Evaluation: video recording, interview + log analysis

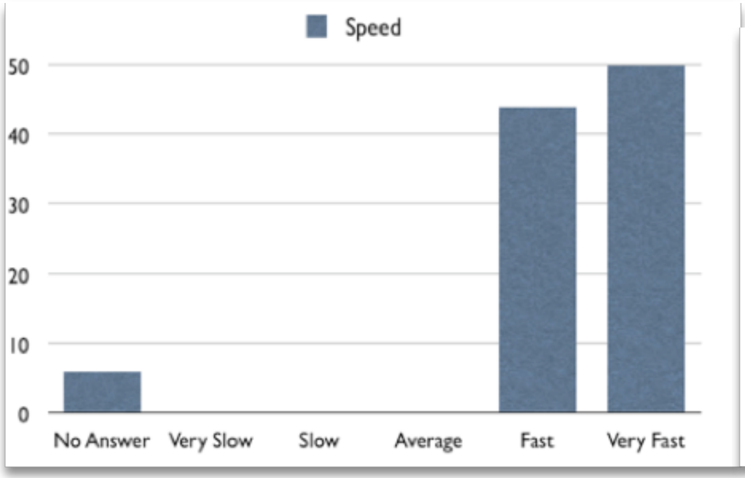
Evaluation Questions

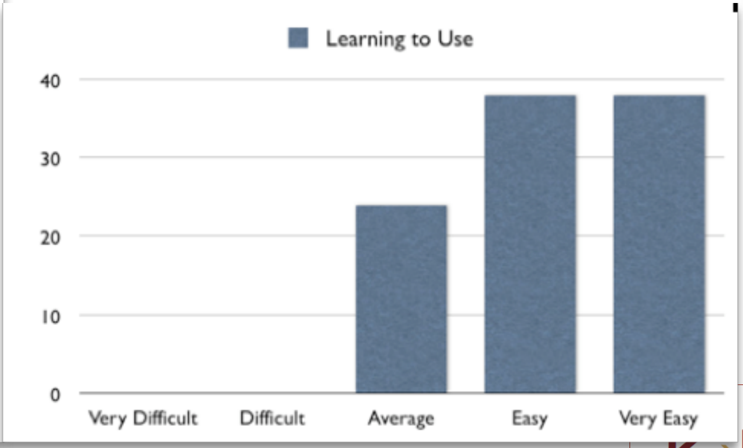
- Do user understand the hybrid paradigm?
- Are they able to search using HS?
- Do they actually use HS when confronted with a real searching task?
- Would the users be willing to use the system for their everyday work?

Liked by the users?









k-now.co.uk

Liked by Users?

- Finalist of Rolls-Royce Director's Creativity Award 2007
  - Voted by employes for its innovation potential

Liked by Users?

- Support to the design of new jet engine
  - Porting to 9 Information Sources
    - 2008-2009
  - Carried out by:
    - 50% University
    - 50% k-now ltd (university spinout-company)

- Funds requested to UK Government for use of K-Tools for use in manufacturing

Conclusions

- Document annotation can be performed at different levels
  - Ontology-based, braindump, document enrichment
- User centred automated ontology-based annotation
  - For trusted self contained documents (e.g. KM)
    - AktiveMedia
- Automated means of capturing knowledge
  - Several Tasks

Conclusions

- Sharing and Reuse
  - We have seen
    - Document Enrichment
    - Semantic Search

Future Work & Challenges

- Multidisciplinary research for automation
  - NLP has strong role, but complemented with other disciplines
  - SE, ML, II, SWS, HCI
- Annotation
  - Beyond the division between user centred and unsupervised
    - Strong HCI strategies
      - Validation of results across documents
        - How can you validate 2M triples produced by large scale annotation?

Future Work & Challenges (2)

- How modelling uncertainty?
- Knowledge is dynamic. How do you model that?
- HCI
    - Information presentation (document annotation)
        - Intrusivity:
            - How to avoid annoying users with too many annotations
        - Trust
            - Who do users trust?
                - Tracing preferred sources
            - Where does the information come from?
- Scalability
    - Large scale indexing systems
        - Millions of pages (not billions!)

Conclusions and Future Work

- 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!

A final thought

- These technologies allow easy collection of *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
  - Just adding hay to the haystack while searching for a needle?
    - Drowning in triples while trying to avoid drowning in texts?

The Karen Spark-Jones slide

Thank You

- Contact Information
  - www.dcs.shef.ac.uk/~fabio
  - fabio@dcs.shef.ac.uk
- Intelligent Web Technologies Lab
  - http://nlp.shef.ac.uk/wig/
- NLP Sheffield
  - http://nlp.shef.ac.uk/
- University of Sheffield
  - www.shef.ac.uk

web intelligence technologies lab

fabio ciravegna

quick links

- my home page
- my research
- publications
- projects
- curriculum vitae
- highlights and news
- teaching
- contacts

news

- eArchaeology: new project funded by AHRC + EPSRC approved! In collaboration with the University of York,

professor of language and knowledge technologies
x-media director

web intelligence technologies lab, .
natural language processing group, .
department of computer science, .
university of sheffield

regent court,
211 portobello street,
sheffield, S1 4DP,
united kingdom
tel:+44(0)114-22.21940,
fax:+44(0)114-22.21810

fabio@dcs.shef.ac.uk

web intelligence technologies

Home | People | Research | Tools | Contact

Natural Language Processing Group, Department of Computer Science, University of Sheffield

welcome

The Web Intelligence Technologies group is part of the NLP Research Group within the Department of Computer Science at the University of Sheffield.

We do research on, and provide applications of, Intelligent Technologies for tomorrow's Web, with a particular focus on their use for Knowledge

document collections;

ies;

this includes:
a particular focus on
notation;
formation in large

and currently

grants news

X-Media
IP project starting in March 2006. X-Media is coordinated by our group. With a budget topping **€13m,** it is one of the largest projects funded in the area of IST knowledge and content technologies within Framework 6 of the European Union.

MultiMatch
a framework 6 strep project on multimedia and multilanguage access to cultural heritages

AHRB
"Information mining in distributive research datasets in the arts and humanities" (£139,000) (June 2005, two years).

DTI-IPAS

nlp sheffield

natural language processing group

home | people | research | life | demos & resources | jobs & study | links | search & sitemap

Fields of research

- NL analysis
  - Biomedical Informatics
  - Information Extraction/Data Mining
  - Dialogue
  - Machine Learning
  - Question Answering/Summarization
  - Reuse and alignment
  - Semantics
  - Word sense disambiguation
- Generation
  - Hypertext Generation
- Resources
  - Corpora
  - Dictionaries

Home

What we do

The Natural Language Processing Research Group is wi
Computer Science at the University of Sheffield.

Home | About | Departments | Contact | Log into MUSE   Search

The University Of Sheffield.

The Achievement Of Excellence.

Study

- Prospective undergraduates
- Prospective postgraduates
- Prospective international students
- Courses and prospectuses

News and Events

University to host top international orchestra
Trevor Pinnock summer residency in Firth Hall
More news releases

University research uncovers earliest British film
Historic discovery at our National Fairground Archive
More news releases

New information for applicants to Medicine and Dentistry
The UK Clinical Aptitude Test (UKCAT)
More on UKCAT

Information

- How we compare
- Research
- Business and industry
- Alumni/Graduates
- Parents
- Schools
- Visitors
- Media Centre
- Jobs
- Staff
- Current students

Quick Links

- Library
- Student Services
- Students' Union
- Making a gift / legacy
- Conferences
- Residences strategy
- The University in the media
- Bursaries
- Graduation 2006

A very Incomplete Bibliography

Semantic Search

- 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

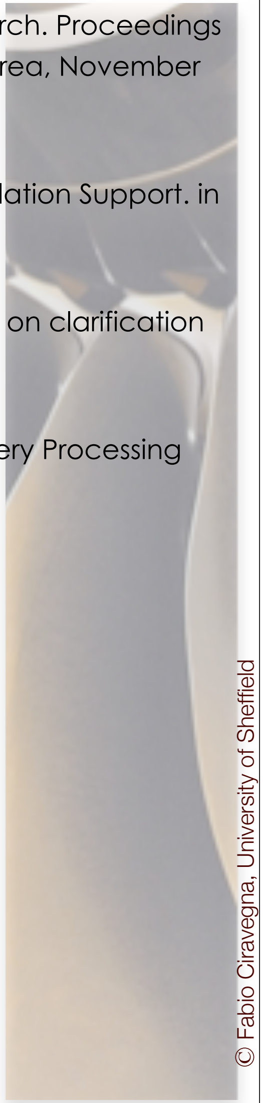
A very Incomplete Bibliography (ctd)

- .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)

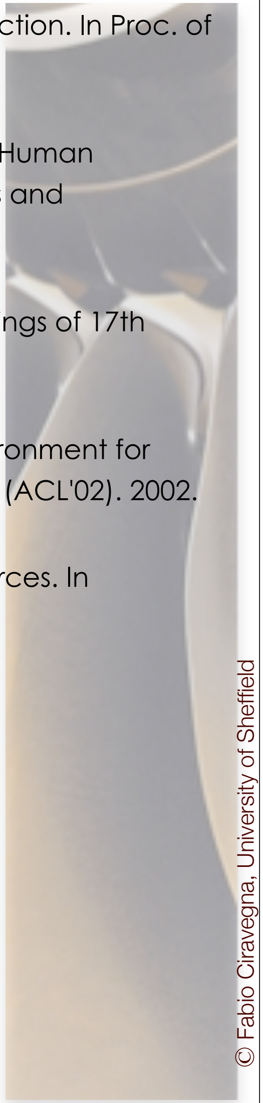


A very Incomplete Bibliography (ctd)

- **Automatic Document Annotation**

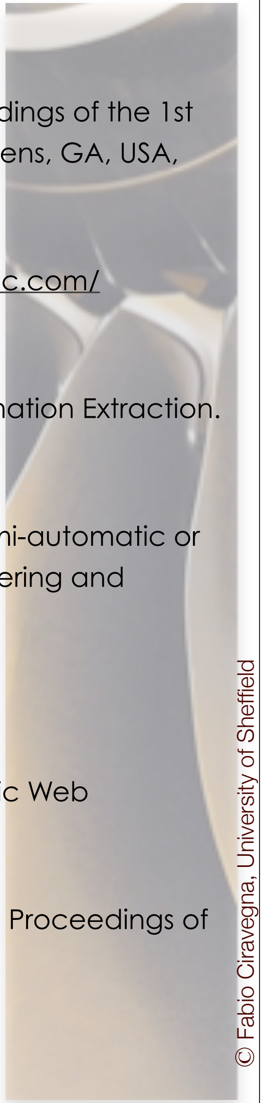
- Fabio Ciravegna. Designing adaptive information extraction for the Semantic Web in Amilcare. In S. Handschuh and S. Staab, editors, Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications. IOS Press, 2003.
- Fabio Ciravegna, Sam Chapman, Alexiei Dingli, and Yorick Wilks: 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

A very Incomplete Bibliography (ctd)



- 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.

A very Incomplete Bibliography (ctd)



Document Annotation

- 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

- F. Ciravegna, A. Dingli, D. Petrelli, Y. Wilks: User-System Cooperation in Document Annotation based on Information Extraction. Knowledge Engineering and Knowledge Management (Ontologies and the Semantic Web), (EKAW02), 2002.

- M. Vargas-Vera, Enrico Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna. MnM: Ontology driven semi-automatic or automatic support for semantic markup. In Proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02. Springer Verlag, 2002

Knowledge Sharing and Reuse

- 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.