Semantic Web Technologies for Capturing, Sharing and Reusing Knowledge

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

The University Of Sheffield.

The Knowledge Life-cycle

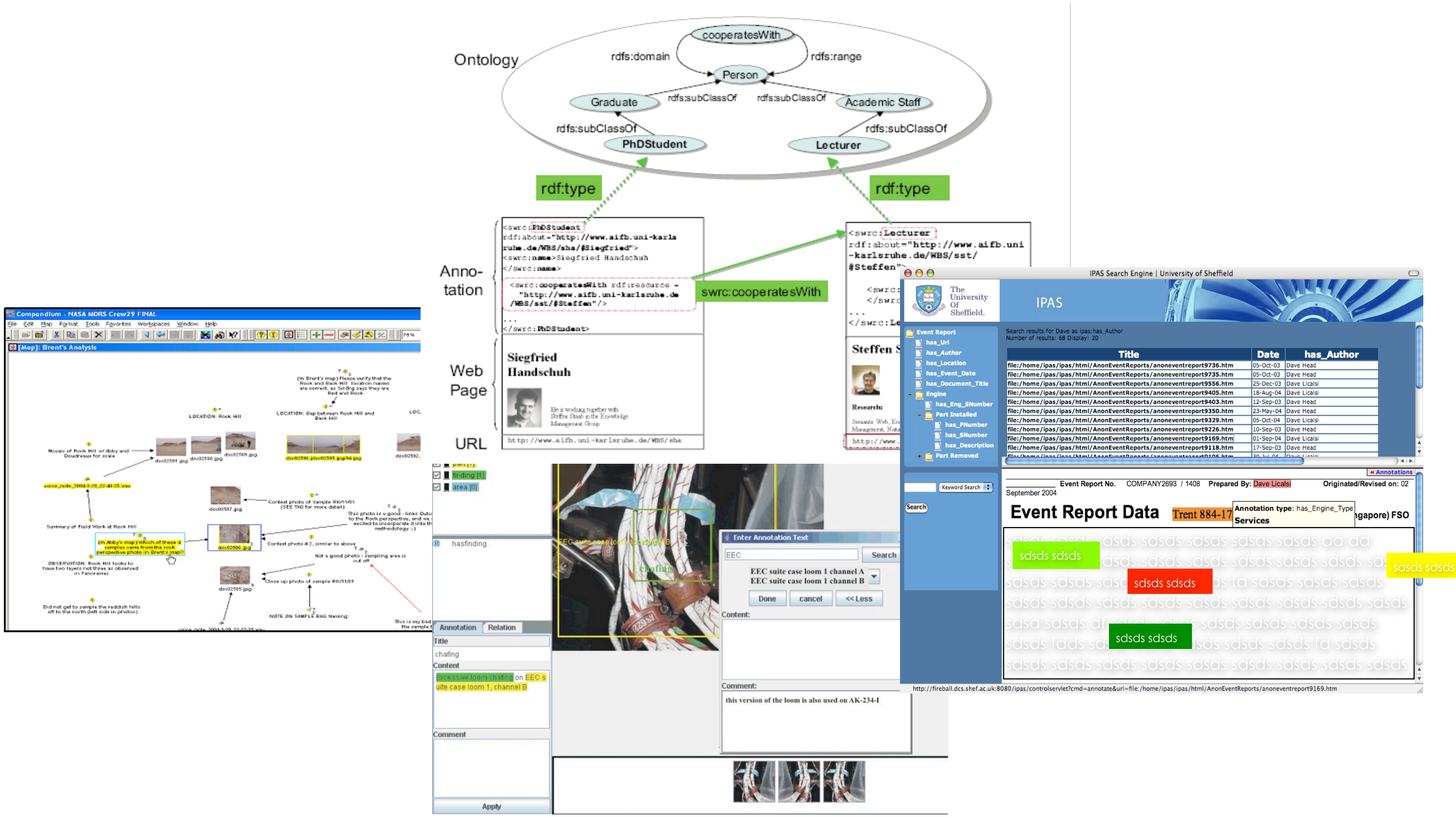
Why manage knowledge?

- 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
- You have already seen:
  - Knowledge Acquisition and Modelling
    - Ontology Engineering



Semantic Web for Knowledge Capture

Knowledge Capture

- Collecting and aggregating knowledge within and across media / archives
  - in a rich, semantically-oriented way
- Two main tasks
  - Annotating existing structured resources
    - E.g. Databases or dictionaries
      - See Guus' talk
  - Annotating unstructured documents
    - Texts, images, data, etc.
    - This tutorial

Semantic Web for KC from Docs

- Two moments to capture knowledge
  - At source: helping people capturing knowledge when produced
  - On legacy documents, pictures, data:
    - Annotation services
- Outcome of capture
  - A semantic representation of (part of) the content
  - Enrichment of multimedia documents
    - with layers of manually or automatically generated annotation

Deliverabl
e D8.2

Compound Documents & KC

- Typical data objects (text, image, raw data)
  - 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

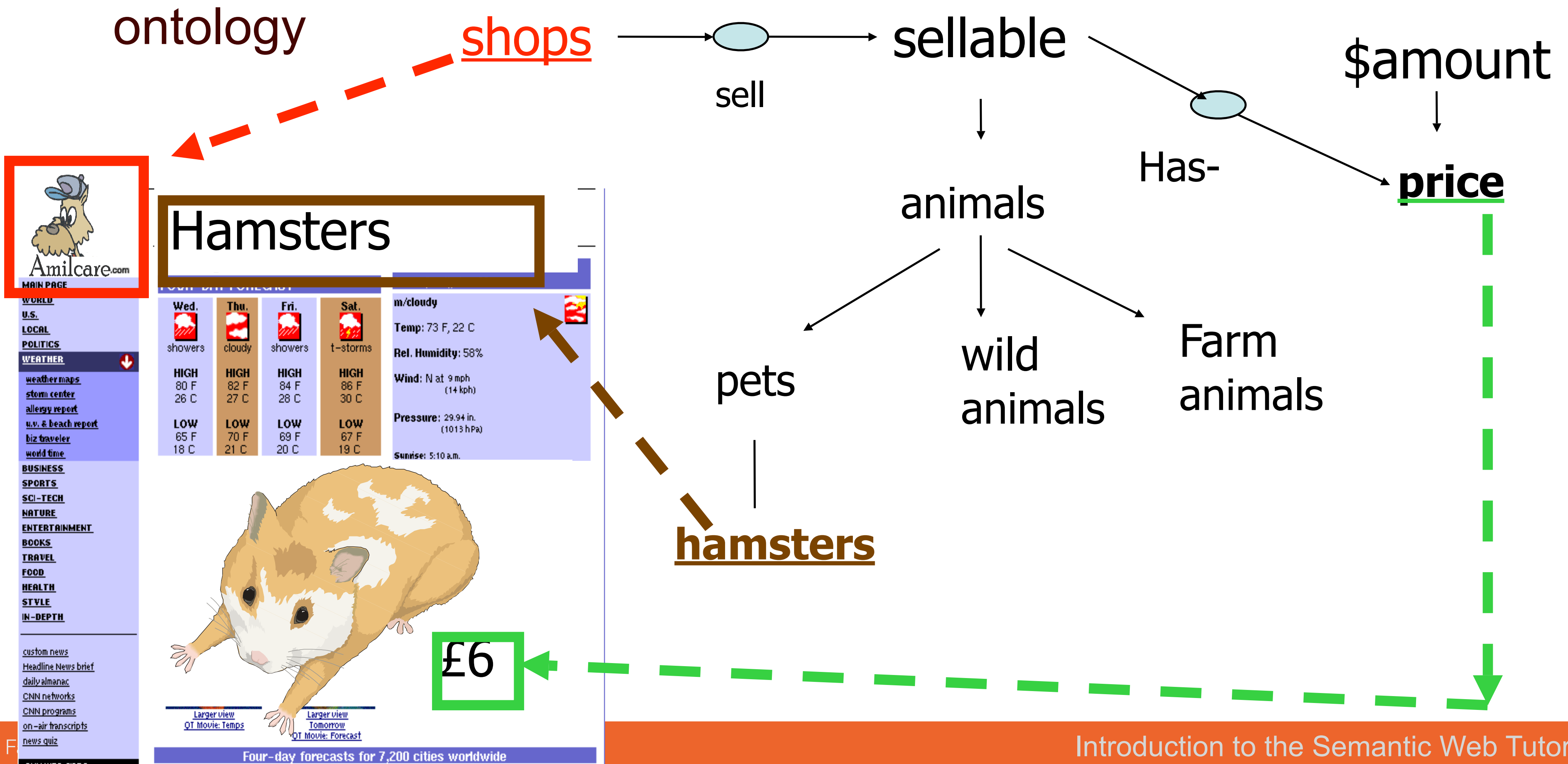
Stripping words off documents takes you nowhere

Ontology-based annotations

- Allows:
  - Ontology-driven processing
  - Services based on ontology will be able to use information

Ontology-based Annotation

- Marking up contained information
  - Portions of documents associated to objects in ontology

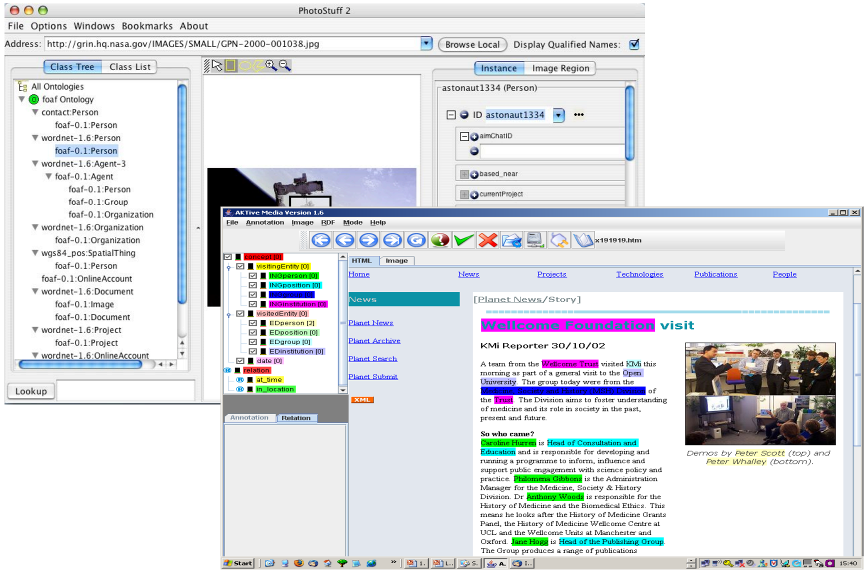


Input &Output

- Input to the KC technologies
  - Ontologies (MMO, domain ontology),
  - Background knowledge (gazetteers, etc.)
  - Documents
    - Possiby in normalised form
  - Medium to extract from (text, images, data, videos,...)
- Output
  - Annotations (e.g. RDF triples stored in triple store)
    - May be in the form of uncertain output

Some Useful Tools

- 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)

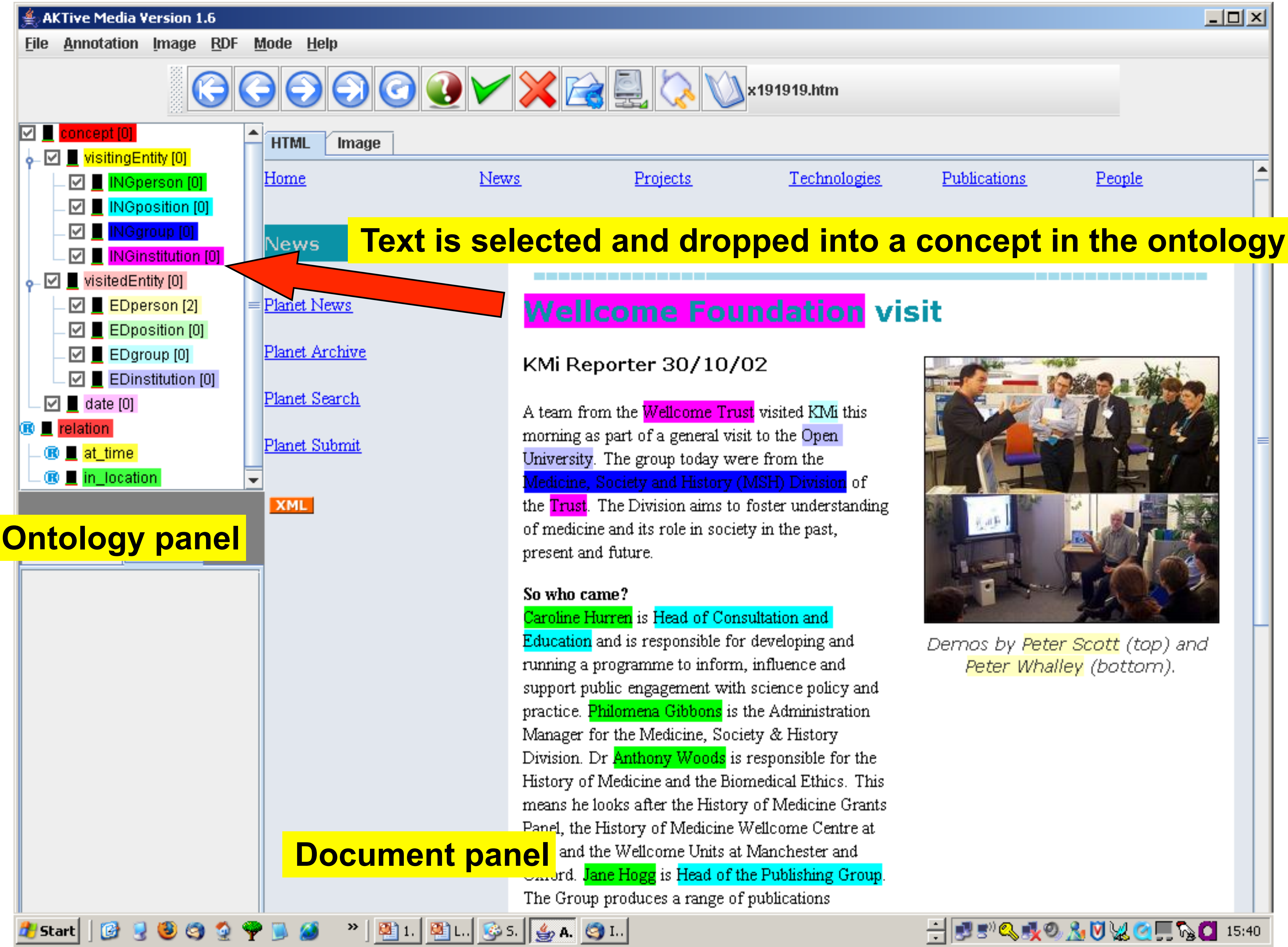


Semantic web 1.0 which was annotating html pages

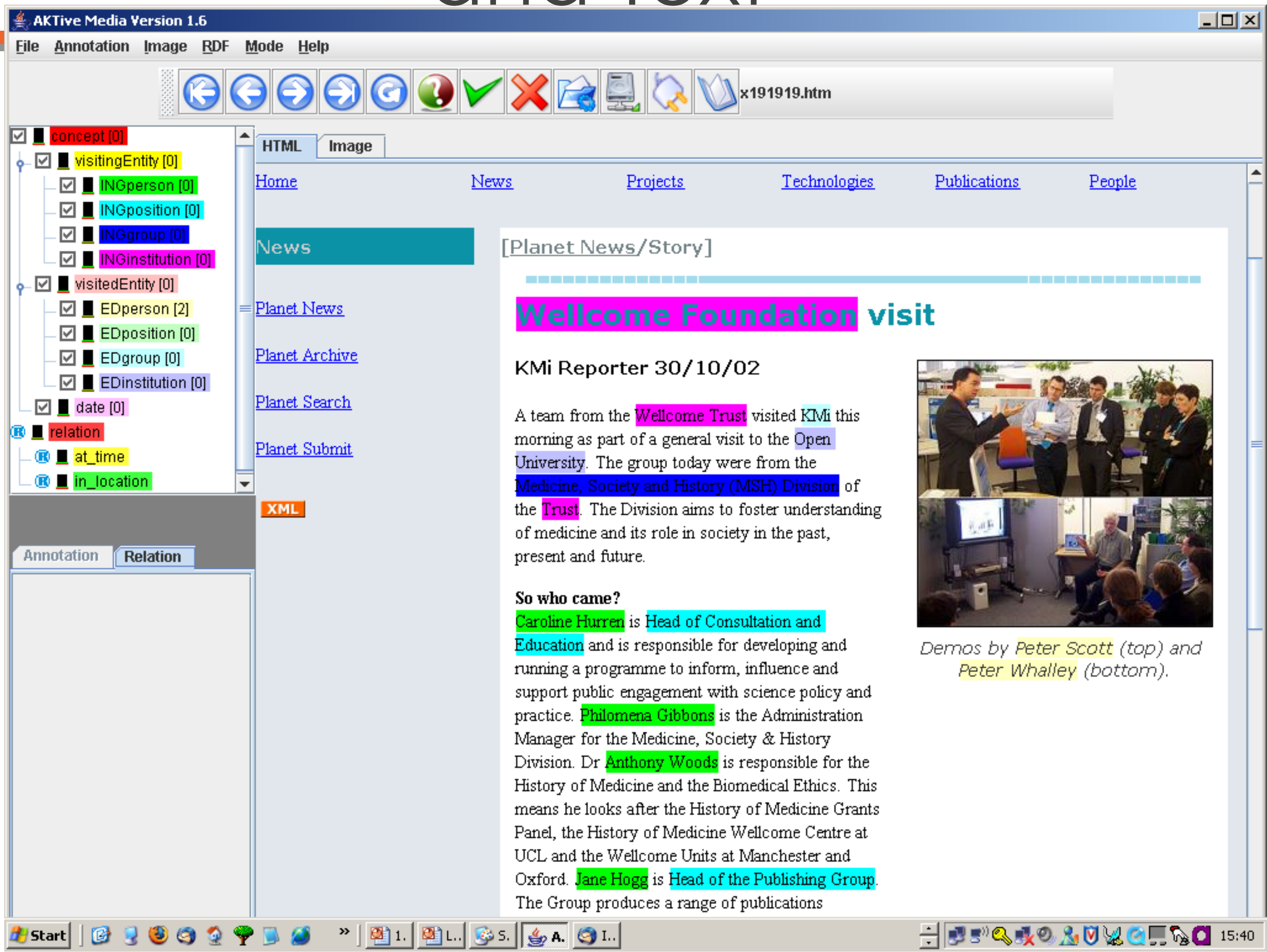
AktiveMedia

- Enables semi-automatic annotation across texts and images
- The interface enables
  - 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

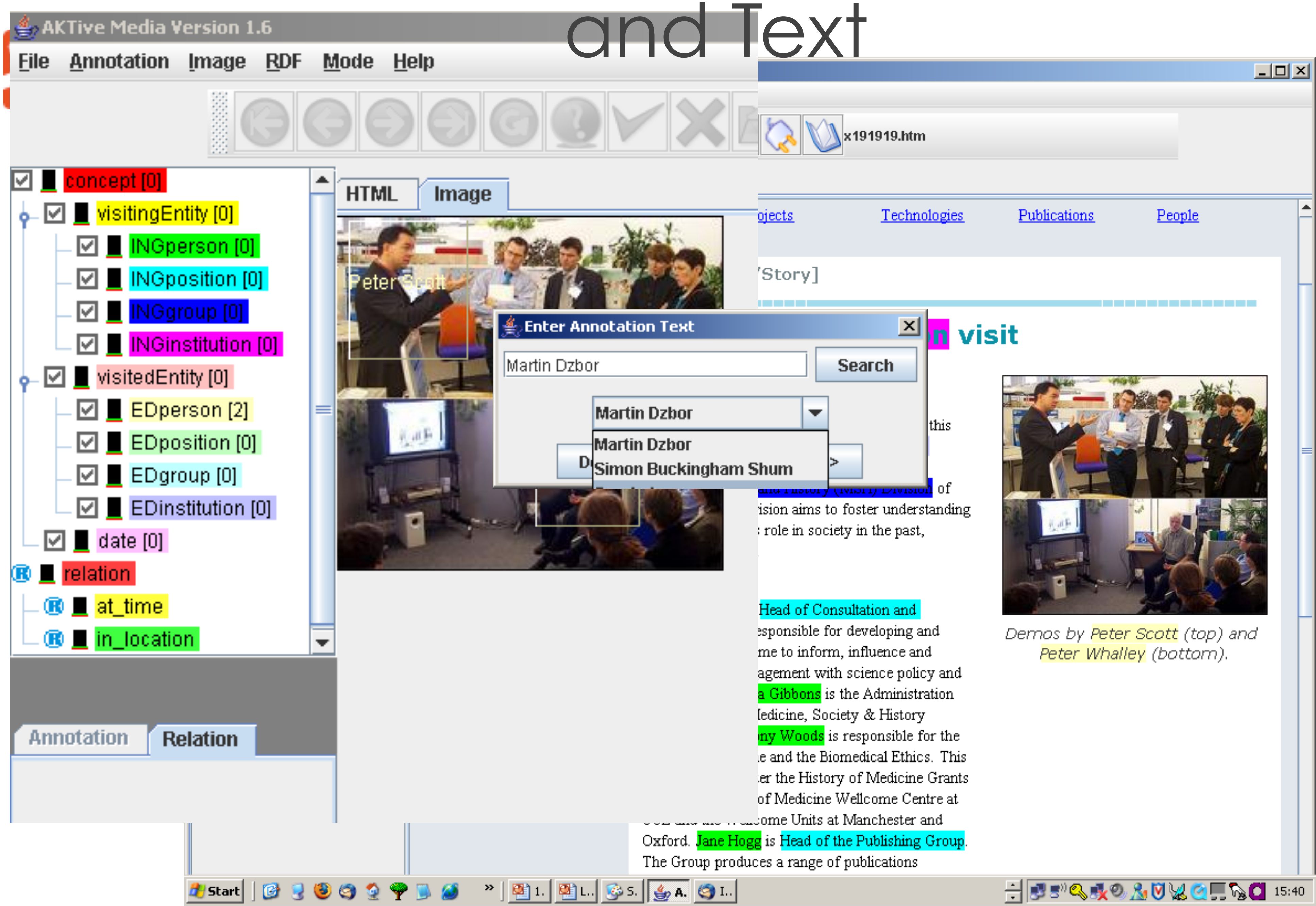
http://www.dcs.shef.ac.uk/~ajay/html/cresearch.html


AKTive Media Version 1.6
File Annotation Image RDF Mode Help
x191919.htm
concept [0]
visitingEntity [0]
INGperson [0]
INGposition [0]
INGgroup [0]
INGinstitution [0]
visitedEntity [0]
EDperson [2]
EDposition [0]
EDgroup [0]
EDinstitution [0]
date [0]
relation
at_time
in_location
HTML
Image
Home News Projects Technologies Publications People
News
Planet News
Planet Archive
Planet Search
Planet Submit
XML
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 …ord. Jane Hogg is Head of the Publishing Group. The Group produces a range of publications
Demos by Peter Scott (top) and Peter Whalley (bottom).
Start
15:40

Contextual Annotation of Images and Text

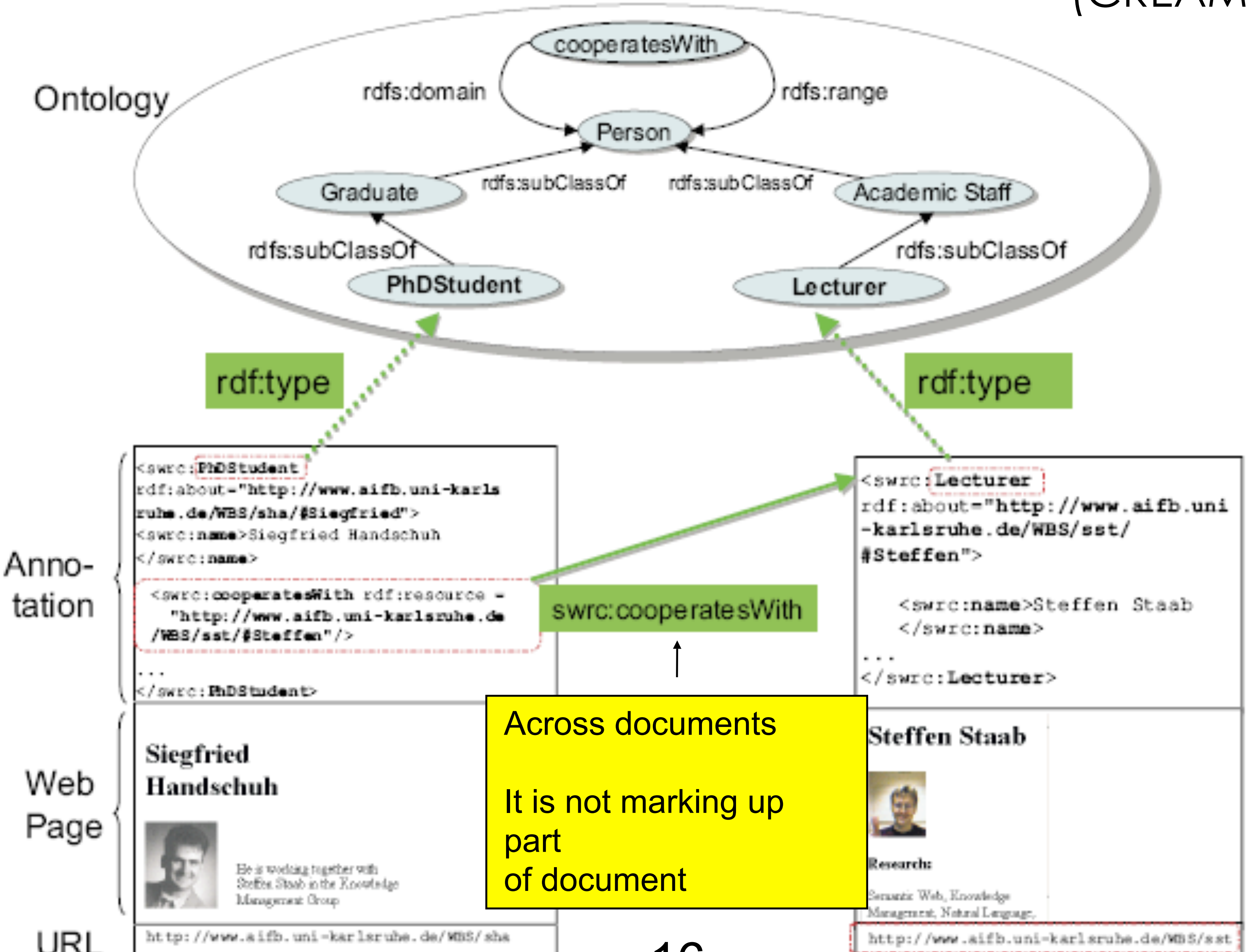


Contextual Annotation of Images and Text



Annotating across documents

(CREAM, 2001)



Marking up Provenance

- COMM - A Core Ontology for Multimedia based on
  - the MPEG-7 standard
  - the DOLCE foundational ontology.

http://comm.semanticweb.org/

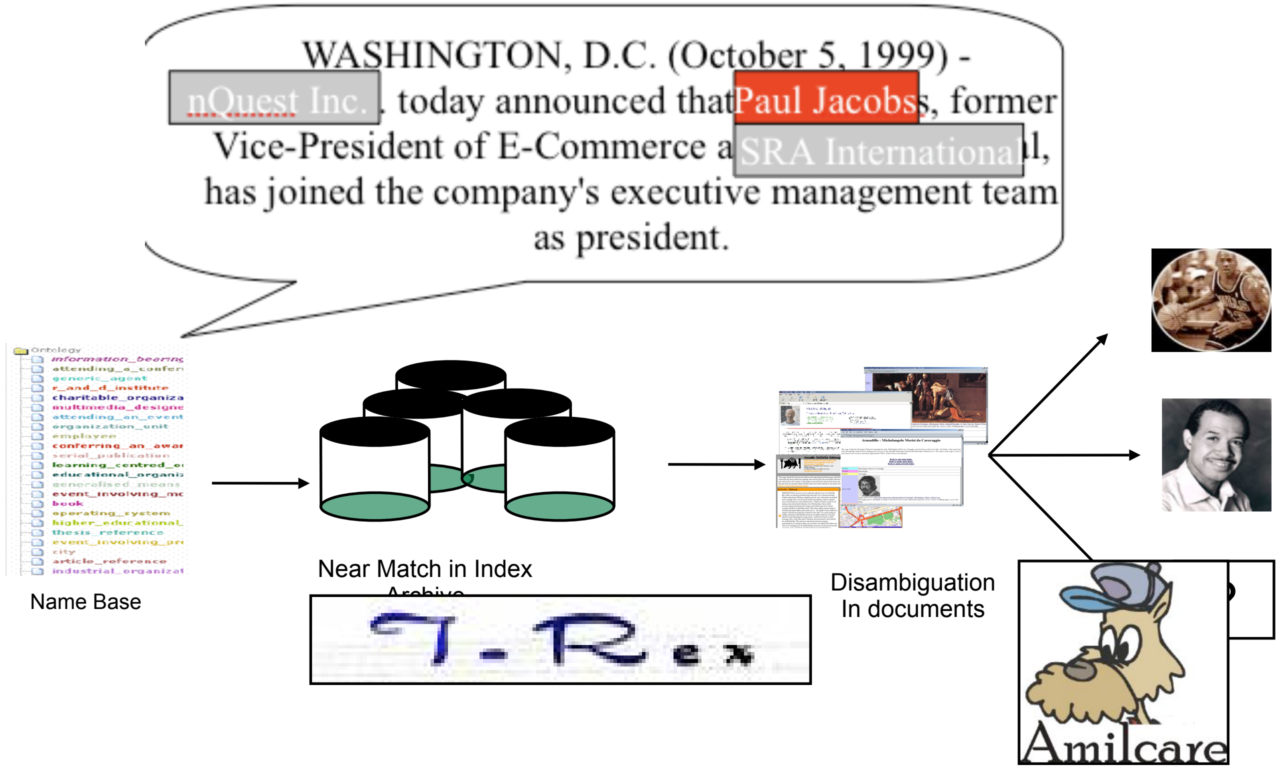
```
<Mpeg7>
 <Description xsi:type="ContentEntityType">
  <MultimediaContent xsi:type="ImageType">
   <Image id="IMG1">
    <SpatialDecomposition>

     <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>
     ...
```

B



Automating Annotation

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
    - Be humble: 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 information source from a semi-structured, unstructured, or free text, information source.

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 information source from a semi-structured, unstructured, or free text, information source.

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

Named Entities — Growing complexity → Event Recognition

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 information source from a semi-structured, unstructured, or free text, information source.

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 information source from a semi-structured, unstructured source.

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

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

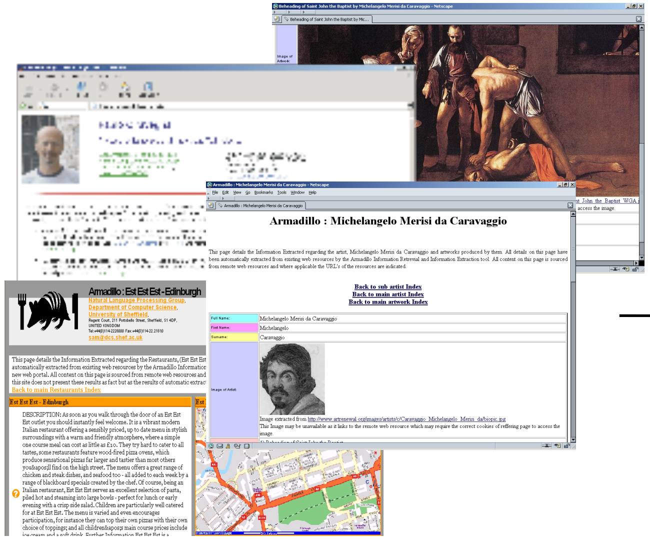
Large Scale NER&C

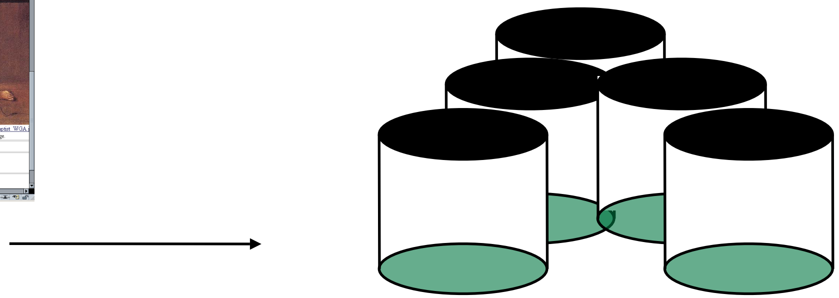
- For large scale (some hundred millions pages) smarter infrastructure is needed
  - Search engine-like indexing infrastructure
  - Faster processing (less processing)
  - Two cases:
    - Recognition of known terms (and their variations)
      - See also information integration
    - Discovery of new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

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

Large Scale NER: Indexing

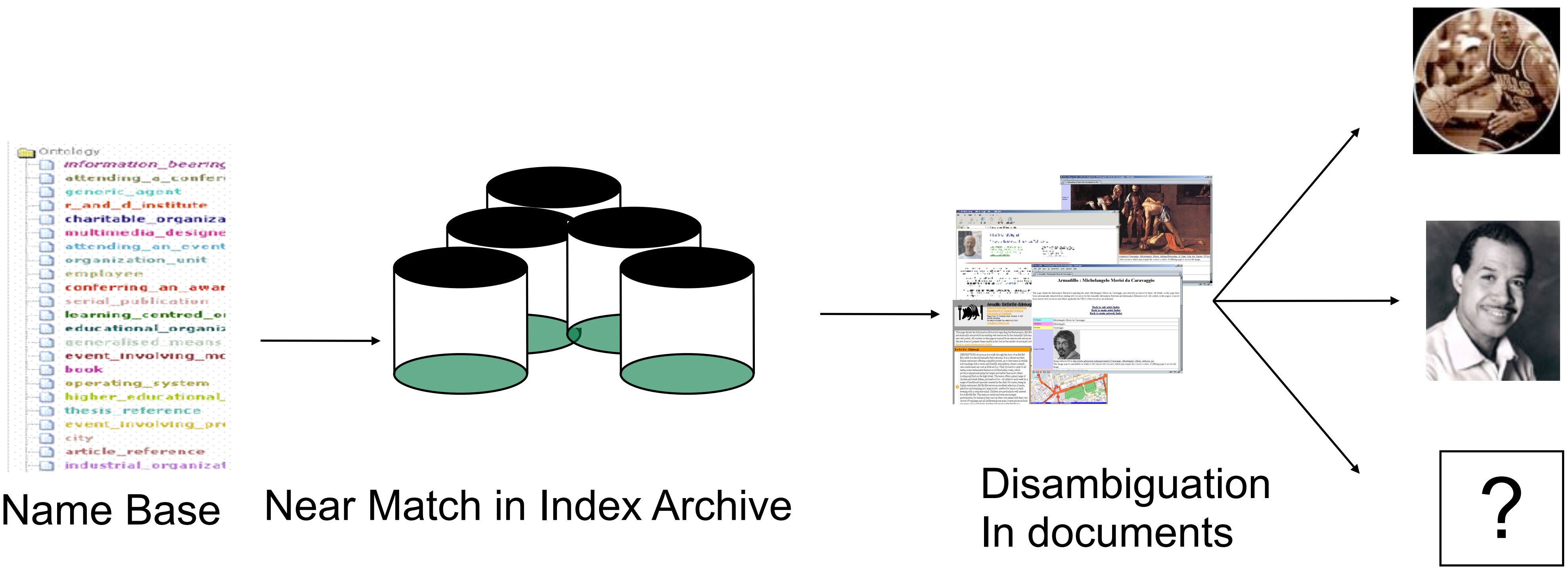
- Document Indexing as in Search Engines





Distributed Index Archive (keywords)

Known Name Recognition



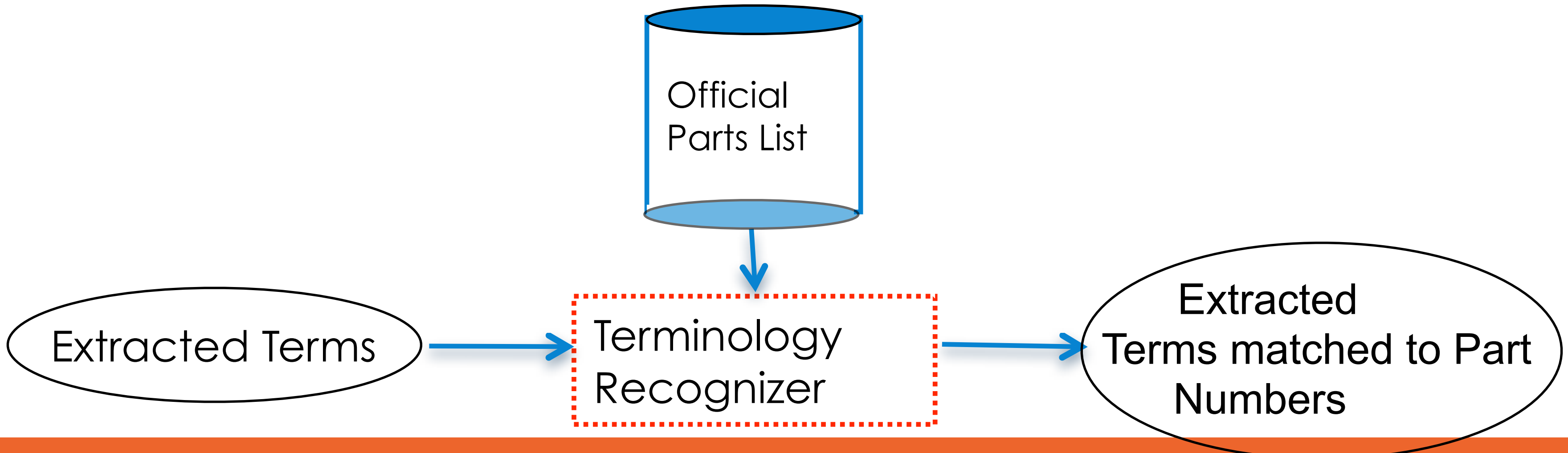
S. Dill, N. Eiron, et al: SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. WWW'03

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



Example of TR

- LP (FAN) COMPRESSOR BLADE
- LP COMPRESSOR SPLITTER FAIRING
- LOW PRESSURE COMPRESSOR BEARING
- HIGH PRESSURE COMPRESSOR BEARING
- SPLITTER FAIRING - LP/IP COMPRESSOR
- 1-6 COMPRESSOR ASSEMBLY
- AIR DUCT - OUTLET IP COMPRESSOR
- BLADE STAGE 6 HP COMPRESSOR

- Query = "Low Pressure Compressor Fairing"

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

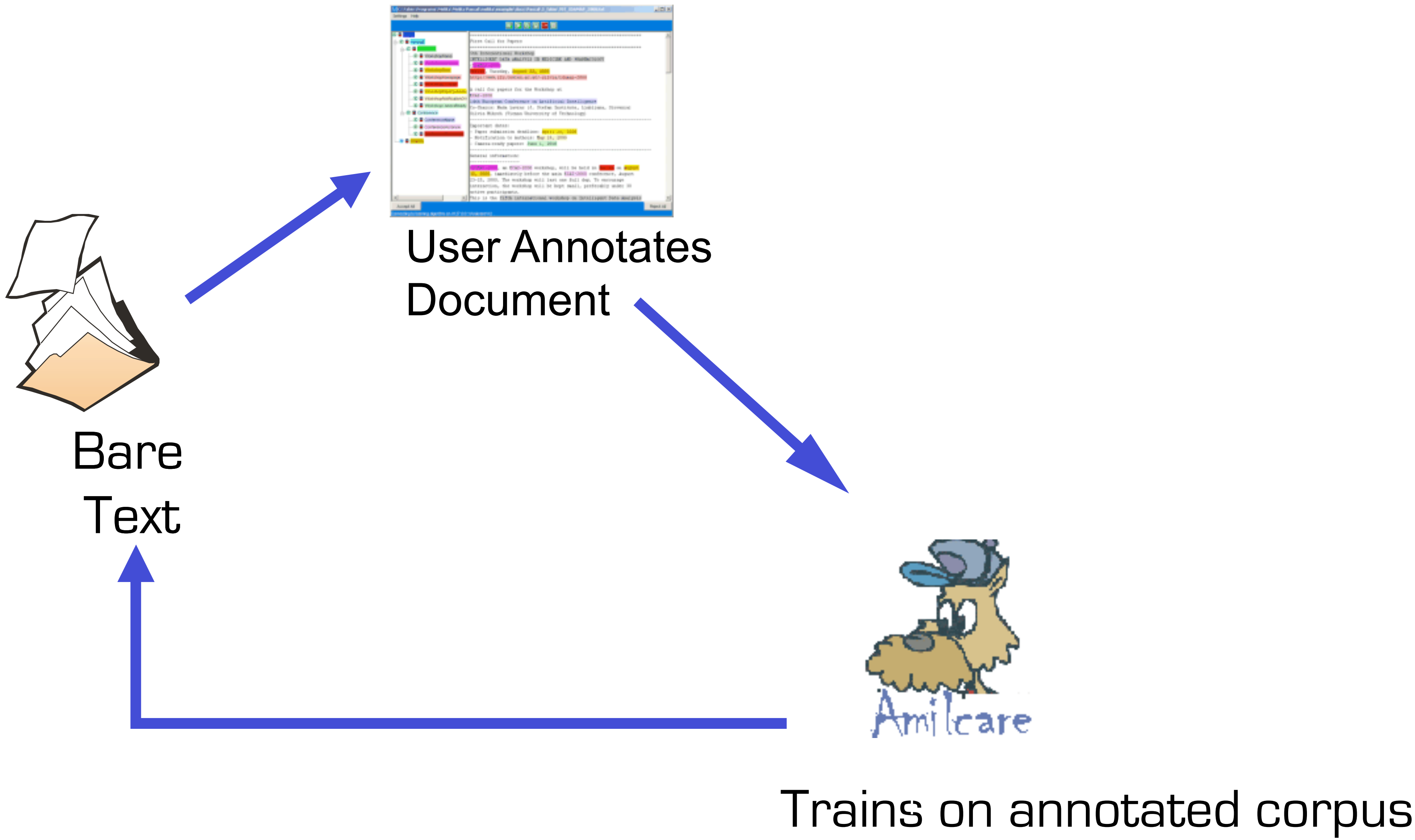
More complex IE: event modelling

- 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

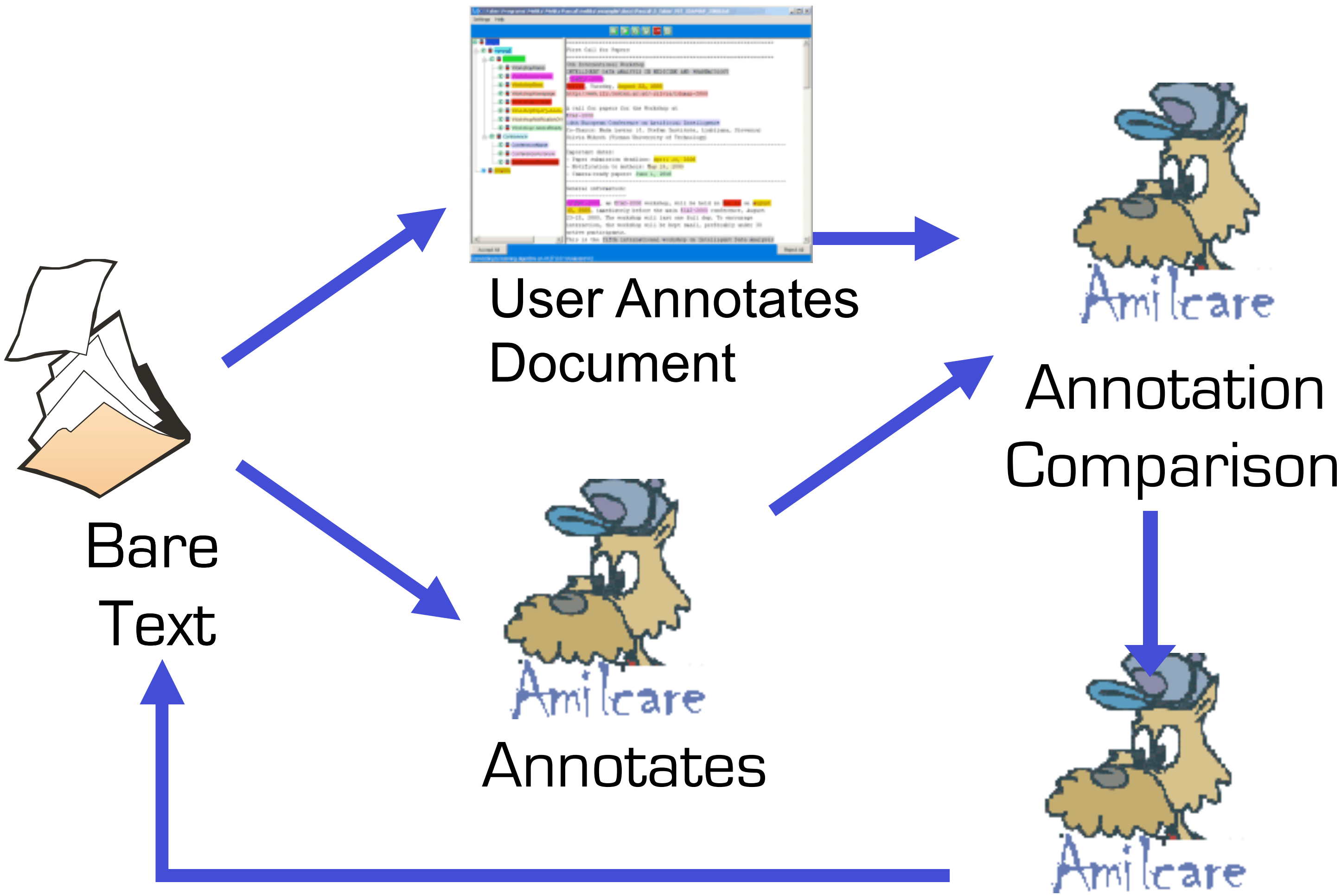
A list of tools for automatic annotation

- 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/

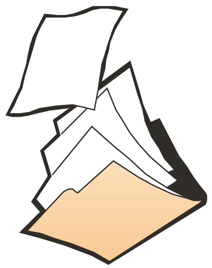
Using IE to support annotation: step 1



Using IE to support annotation: step 1



Using IE to support annotation: step 2



Bare Text

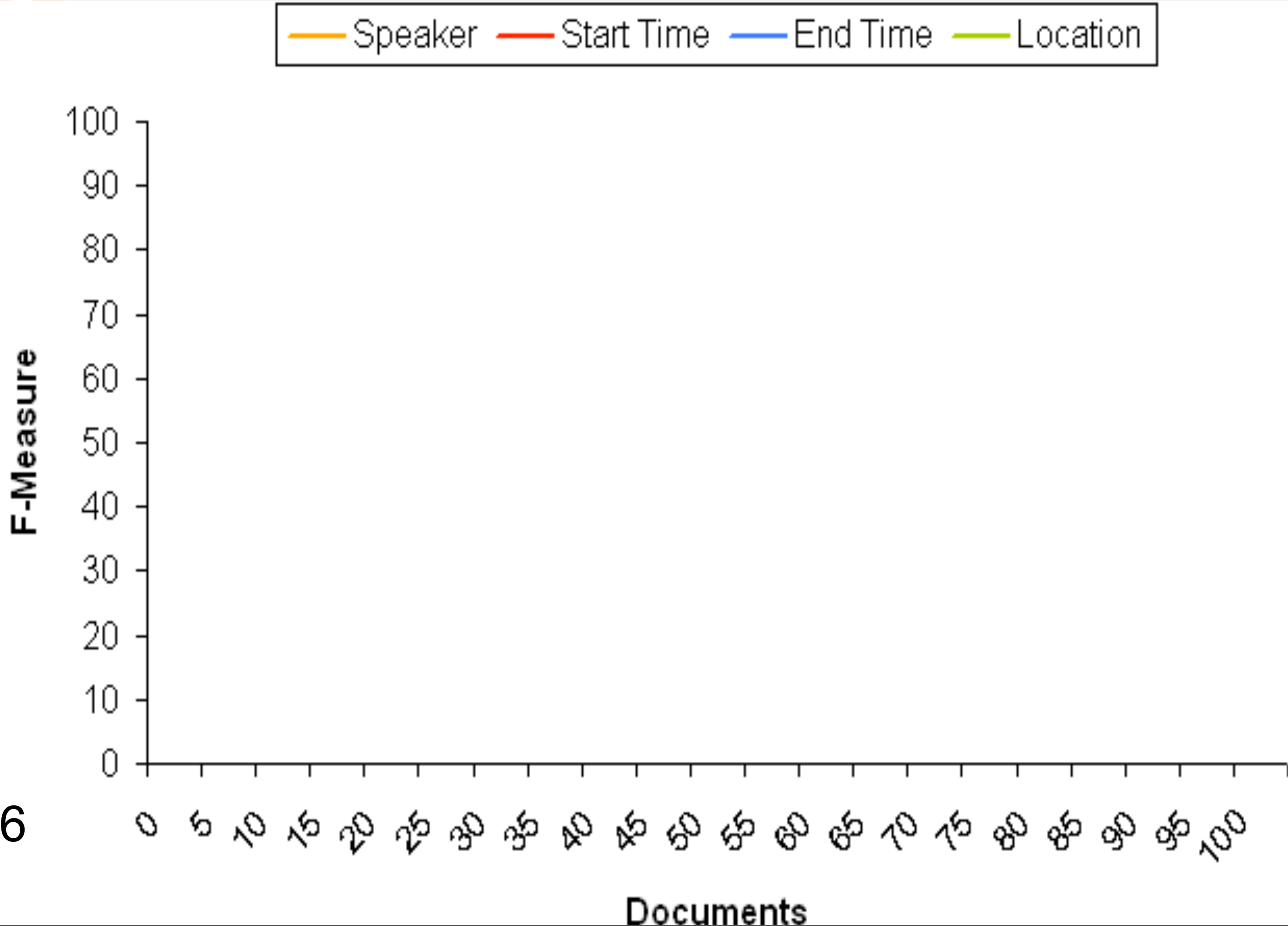
Annotates

User Corrects

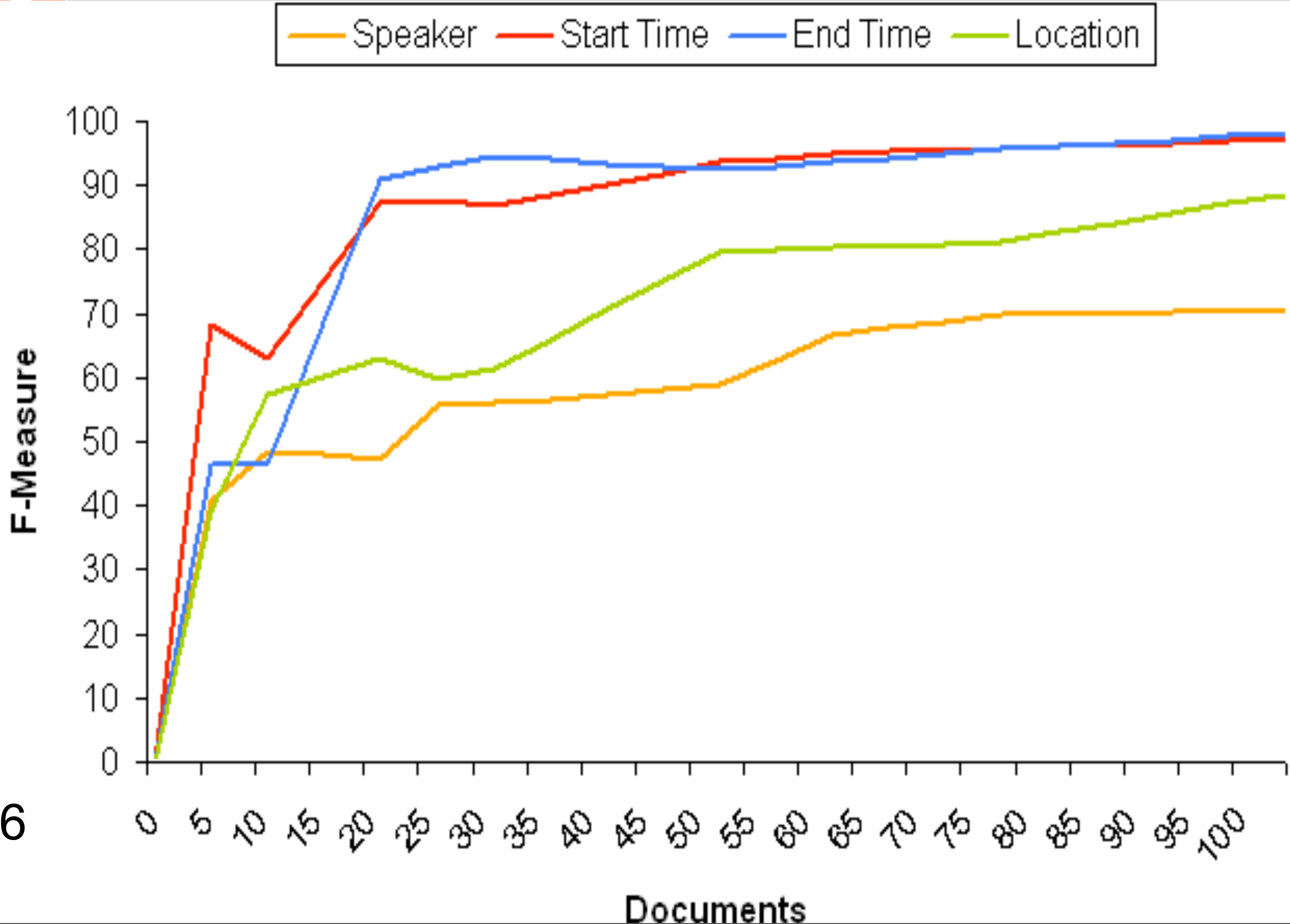
Uses corrections to retrain

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

Learning curve

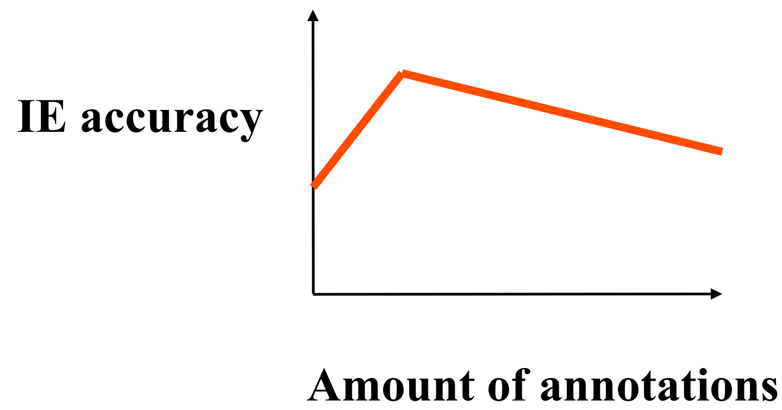


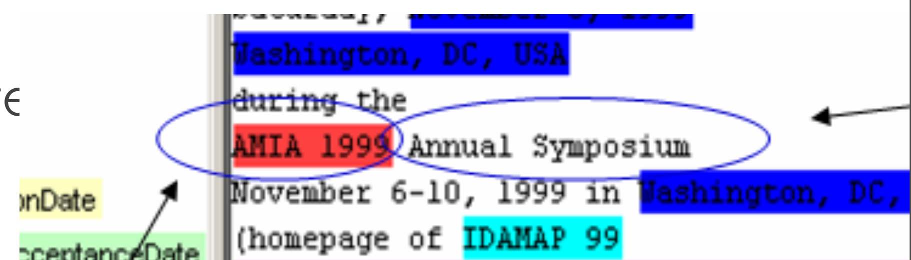
Learning curve



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 consiste
    - Otherwise the IE system is confused
    - Can only annotate document content
      - With connections to the rest of the knowledge via information integration





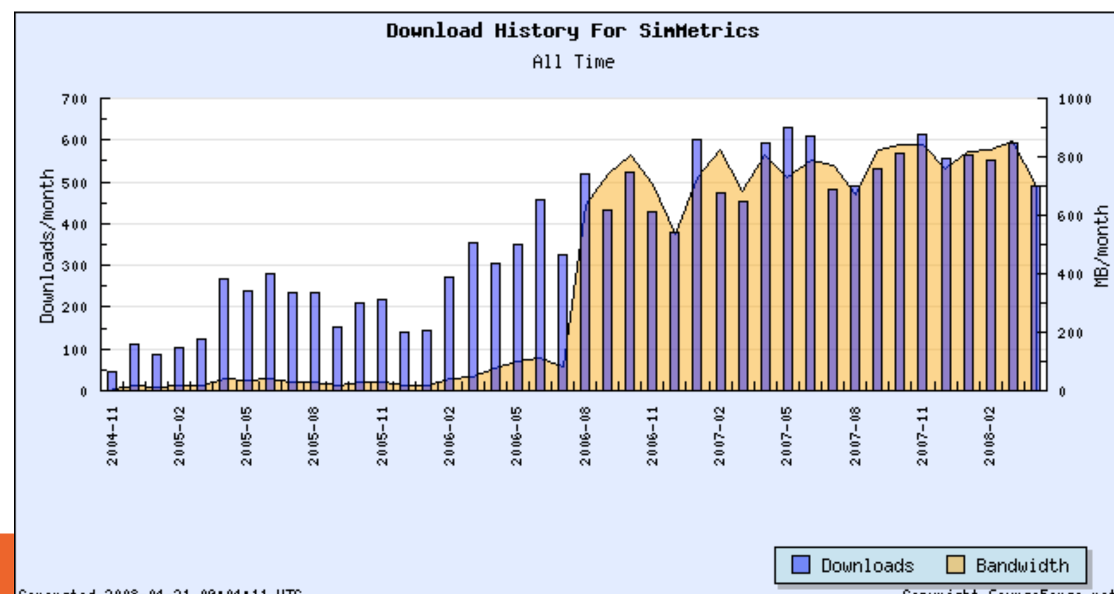
Information Integration

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



Arts & Humanities Research Council

Armadillo: Historical Data Mining

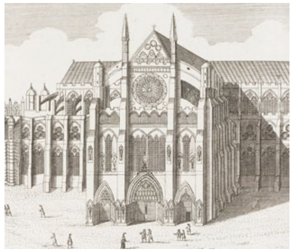
AHDS Deposits

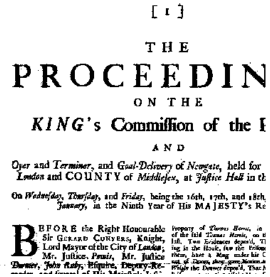
- The Marine Society Registers
- The Westminster Historical Database
- Eighteenth Century Fire Insurance Policies
- Prerogative Court of Canterbury Wills
- The Proceedings of the Old Bailey

- St. Martin's Settlement Exams Index — **WESTCAT**
- Collage image databse — **Guildhall Library**
- Metropolitan London in the 1690s — **IHR**
- Selected Criminal Records — **PRO**
- House of Lords Journals — **BOPCRIS**

- Harben's Dictionary of London
- John Strype's "Survey..."

http://www.motco.com




[1]

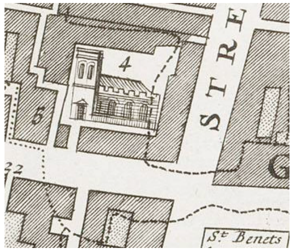
THE

PROCEEDIN

ON THE

KING's Commiffion of the I

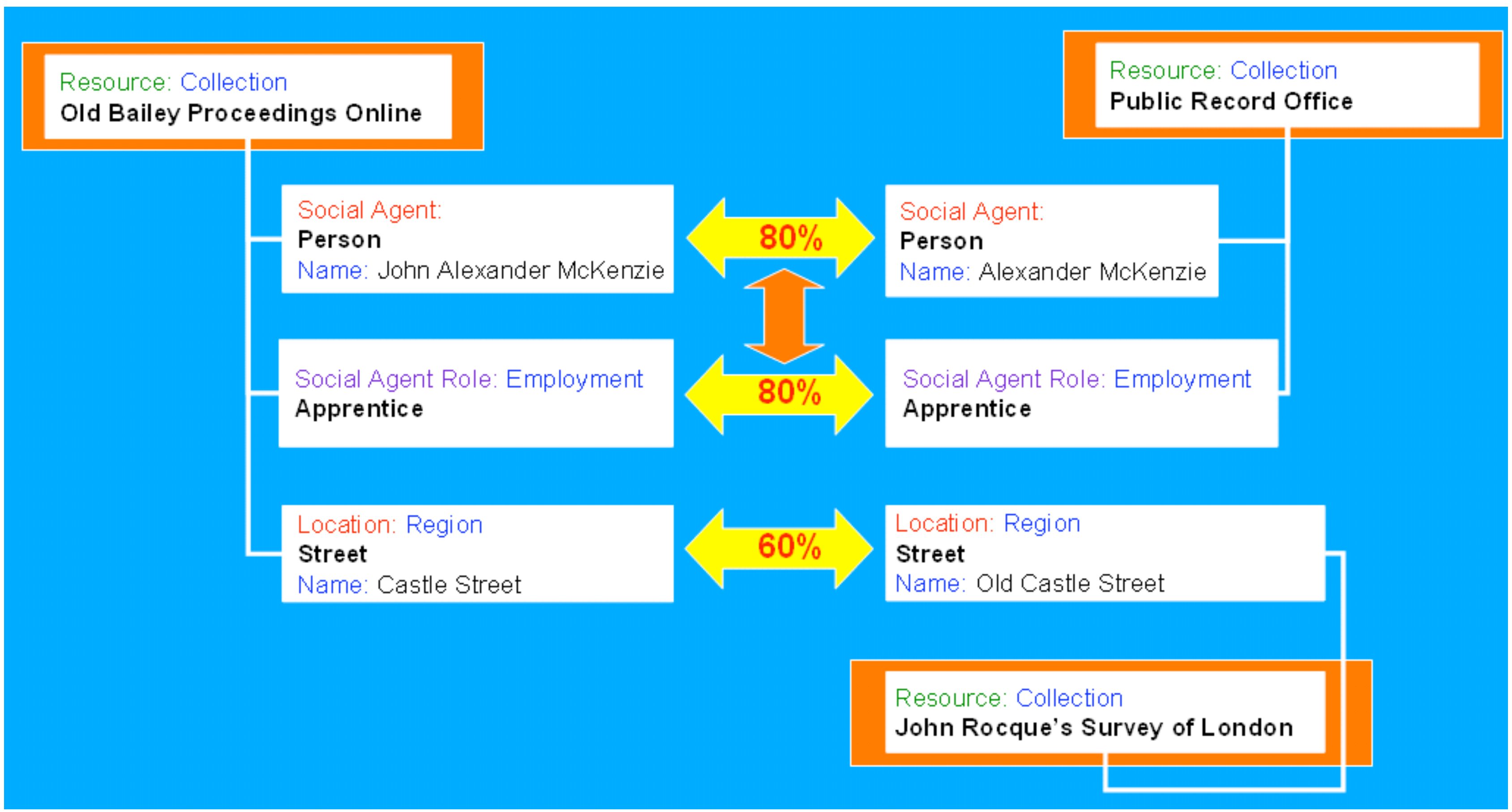
AND



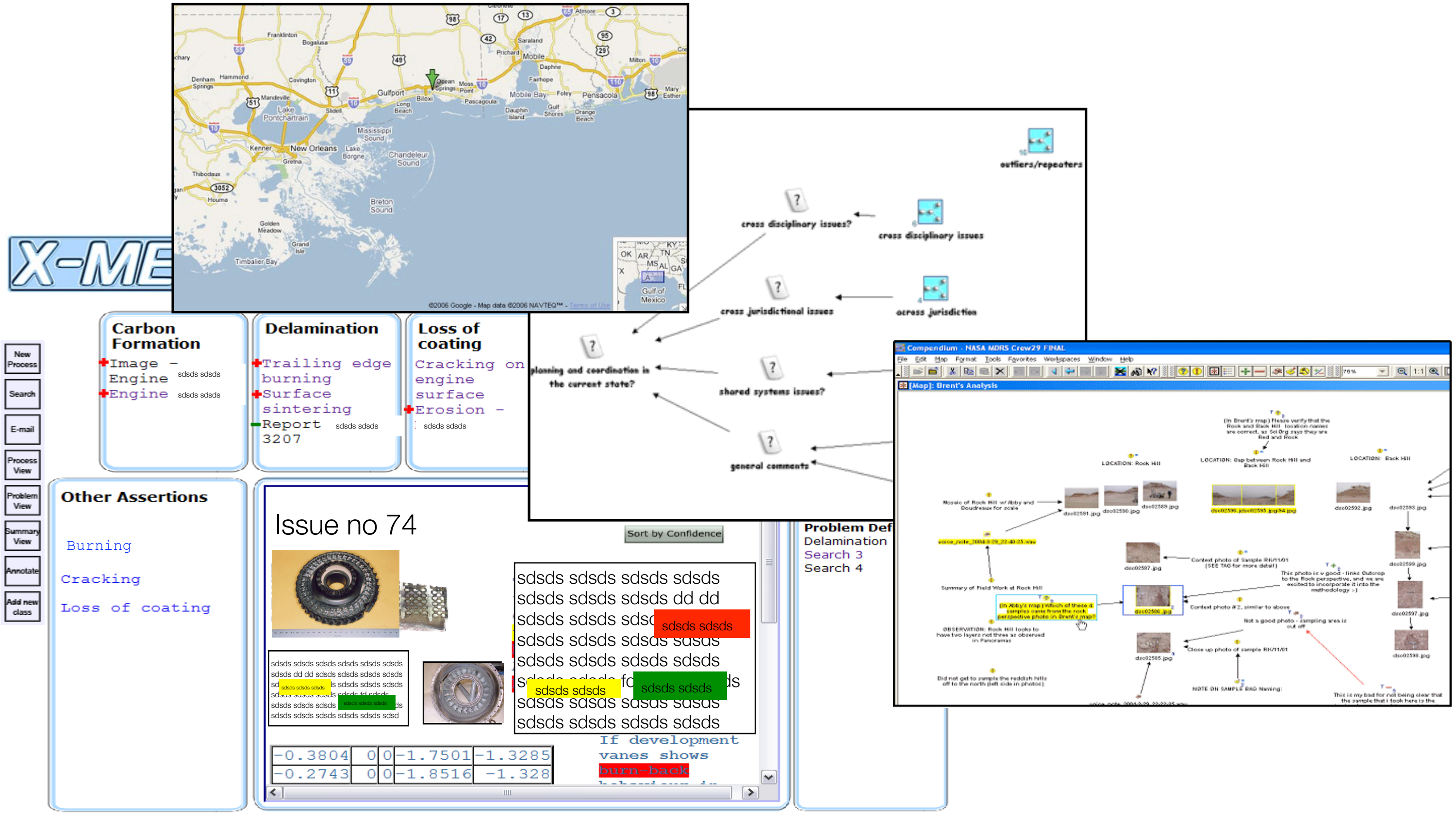
http://www.hrionline.ac.uk/armadillo/

Information Integration

Armadillo: Historical Data Mining



Arts & Humanities Research Council



Sem Web for Knowledge Sharing and Reuse

Knowledge Sharing and Reuse

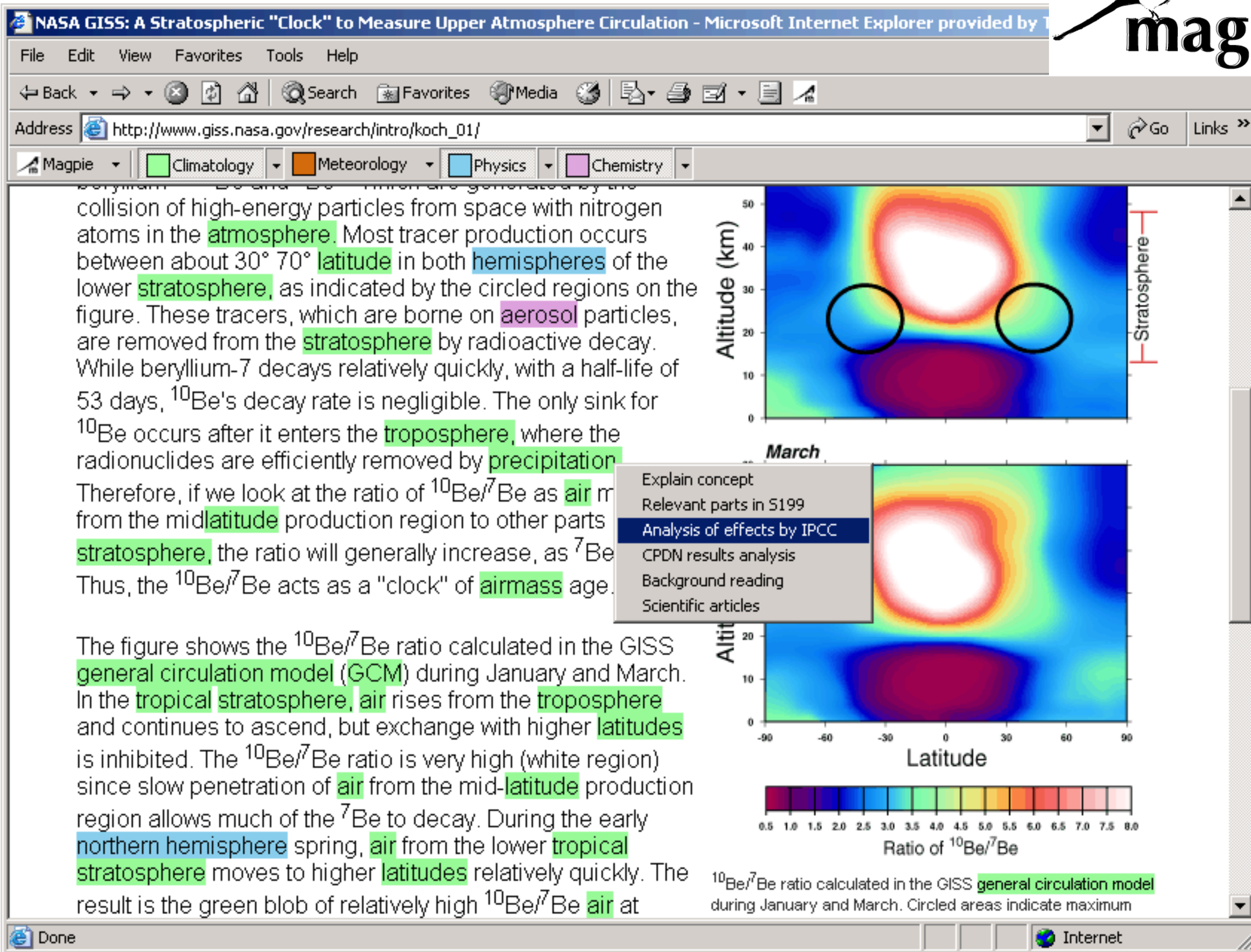
- In KM mainly means
  - Retrieving information and knowledge
    - At the right time
    - In the right form
      - » E.g. independently from where it is stored
      - » Or even the form in which it is stored
    - Suitable to the specific users
      - » e.g. patients should net receive information 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

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
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
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, 10Be's decay rate is negligible. The only sink for 10Be occurs after it enters the troposphere, where the radionuclides are efficiently removed by precipitation.
Therefore, if we look at the ratio of 10Be/7Be as air m
from the midlatitude production region to other parts
stratosphere, the ratio will generally increase, as 7Be
Thus, the 10Be/7Be acts as a "clock" of airmass age.
The figure shows the 10Be/7Be 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 10Be/7Be ratio is very high (white region) since slow penetration of air from the mid-latitude production region allows much of the 7Be 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 10Be/7Be air at
Explain concept
Relevant parts in S199
Analysis of effects by IPCC
CPDN results analysis
Background reading
Scientific articles
Altitude (km)
Stratosphere
March
Latitude
Ratio of 10Be/7Be
10Be/7Be ratio calculated in the GISS general circulation model during January and March. Circled areas indicate maximum
Done
Internet

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

Why Search?

- Task in Searching
  - Document Searching (images/texts/videos/data)
    - Goal of query is to retrieve documents
      - Semantic Search is used as replacement for traditional keyword based systems
  - Knowledge Searching
    - Goal is to retrieve knowledge (i.e. triples)
      - This is similar to search a virtual database
      - Independently from source
        - » Documents can be accessed for
        - » Provenance analysis (checking correctness)
        - » Further browsing

Keywords Based Search

- Document search:
  - 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)
- Knowledge search
  - Not supported

Semantic Search (OS)

- Search instances in an ontology
  - Annotations are unambiguous
    - OS Does not suffer from ambiguity and synonym issues of keyword-based systems (KS)
- Supports knowledge search
  - Naturally
- Supports document search
  - But…

User Centred Approaches

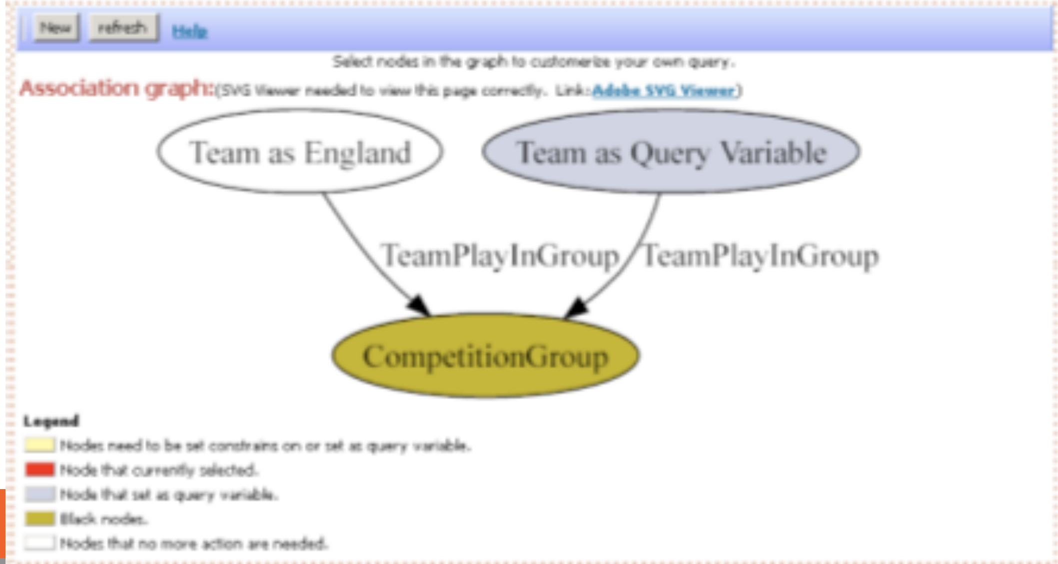
- By merging the definitions in [Uren et al. 2008], [Kaufmann et al. 2007b] and [Baghdev et al. 2008]:
  - Keyword-based approaches considering a natural language query as a bag of words
    - [Kaufmann et al. 2007a] [Lei et al., 2006])
  - Natural language approaches: modelling the linguistics of the 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?
    - What if my keyword does not map to ontology?
- E.g. SemSearch

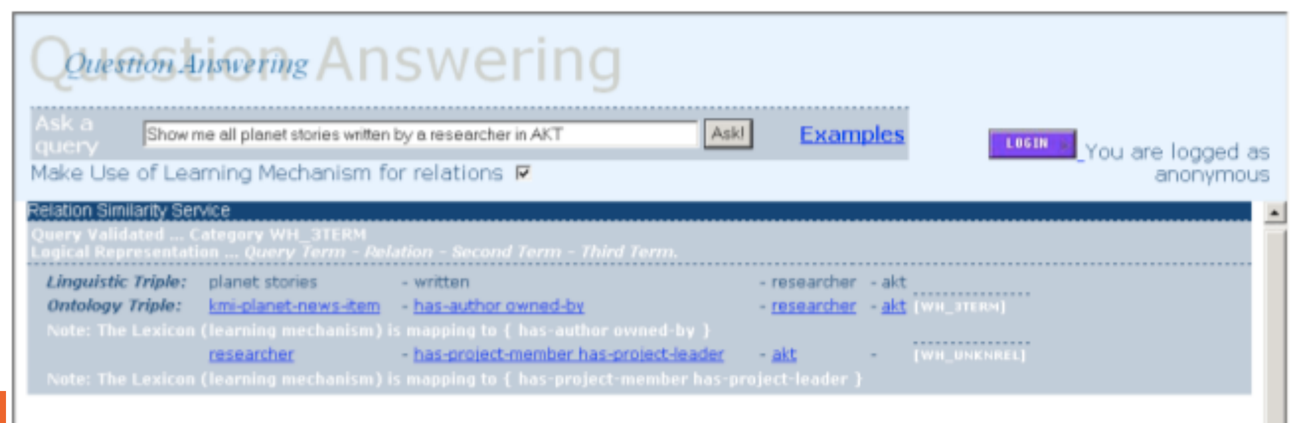
Semantic Search Approaches (2)

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



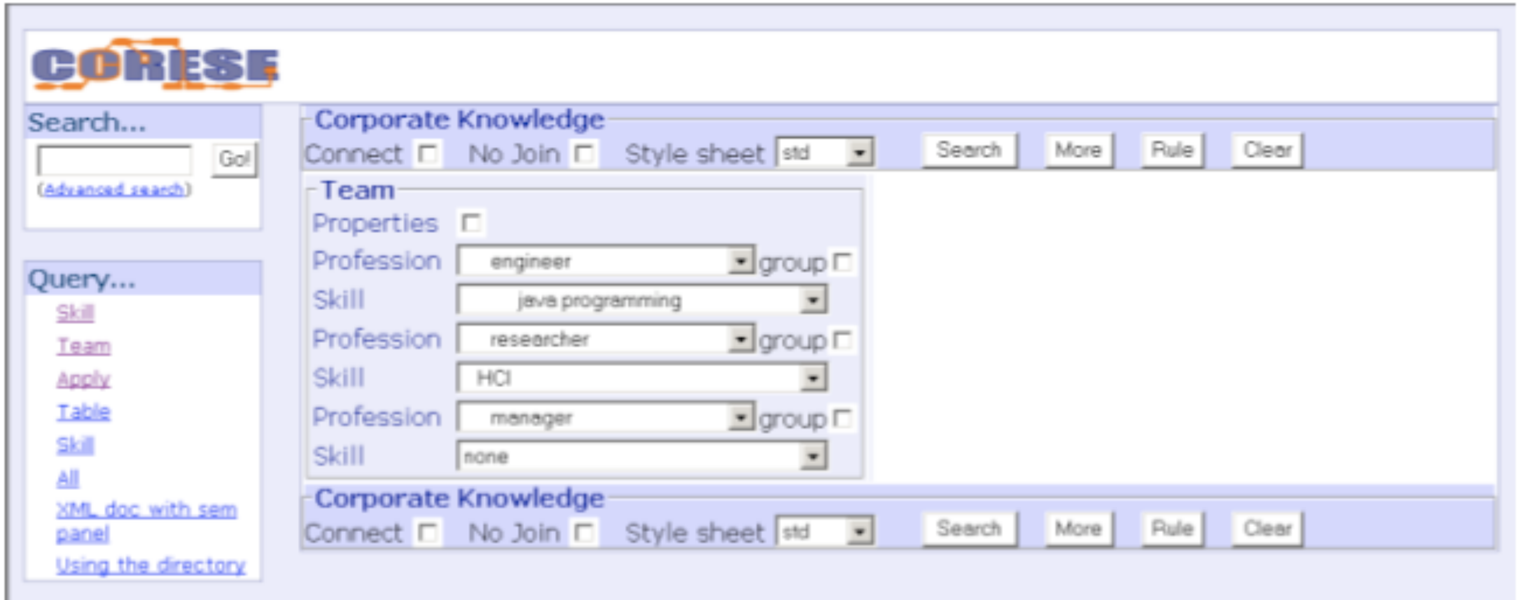
Semantic Search Approaches (3)

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



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



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 search for the parts not 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

- Any boolean combination of three types of 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
    - matching keywords only within portion of documents semantically annotated with a specific type or instance

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

Queries in Hybrid Search

- Any boolean combination of three types of 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
    - matching keywords only within portion of documents semantically annotated with a specific type or instance

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

- 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

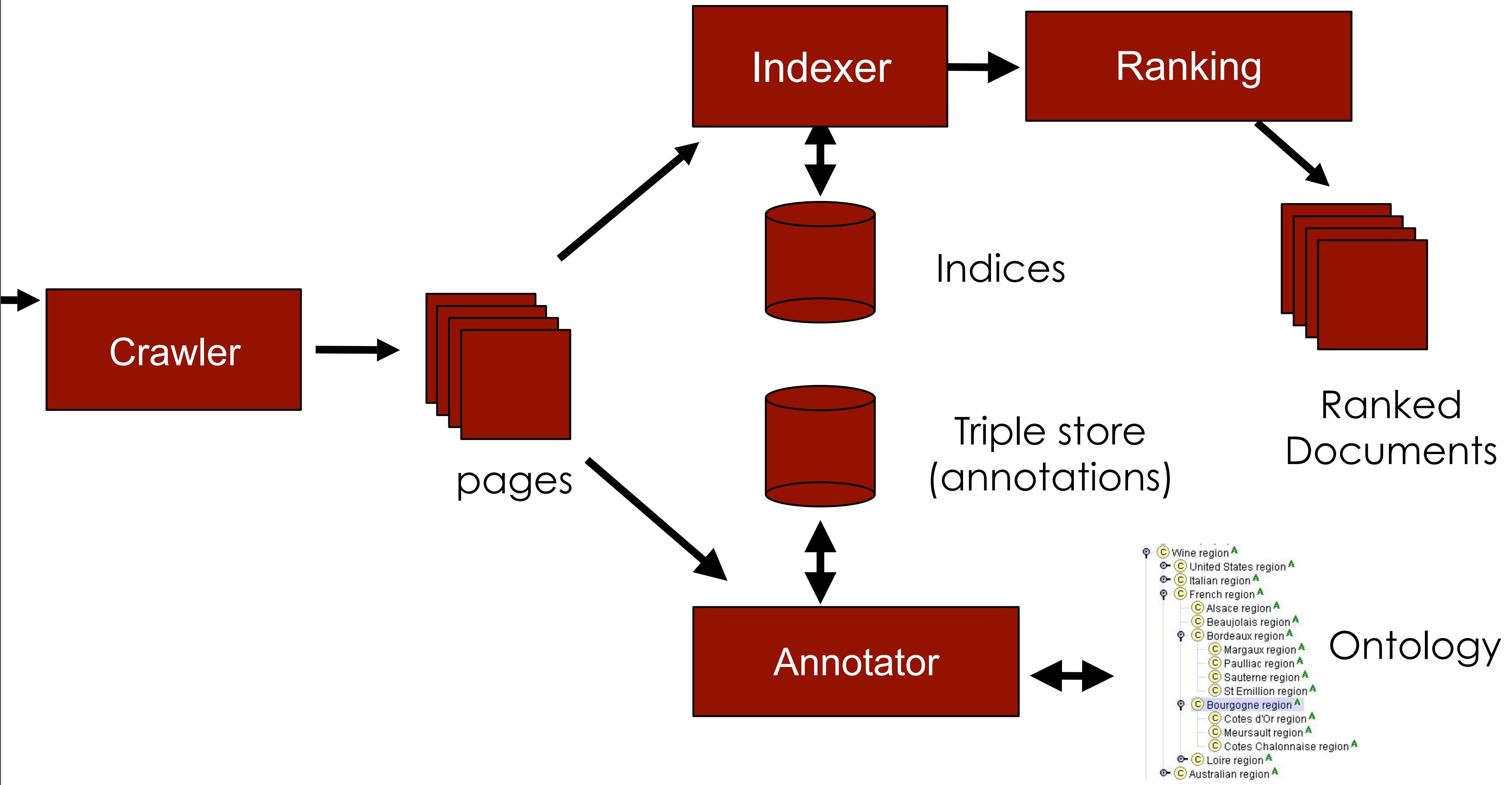
& (provenance-text-contains x “blade”)

Keyword in Context Query

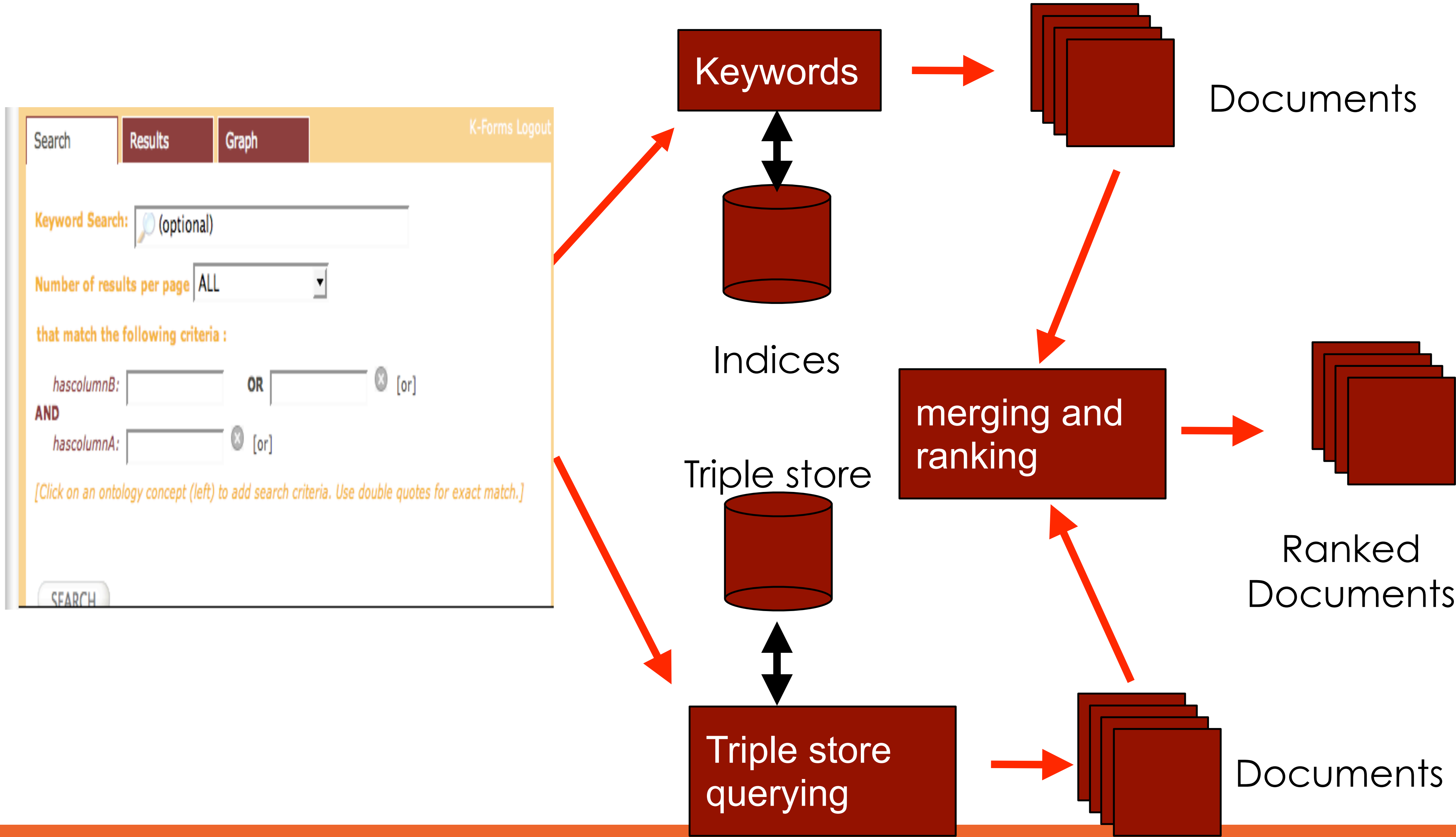
& (contains z “trailing edge”) & (document z) & (provenance x z)

Keyword-based Query

K-Search: indexing



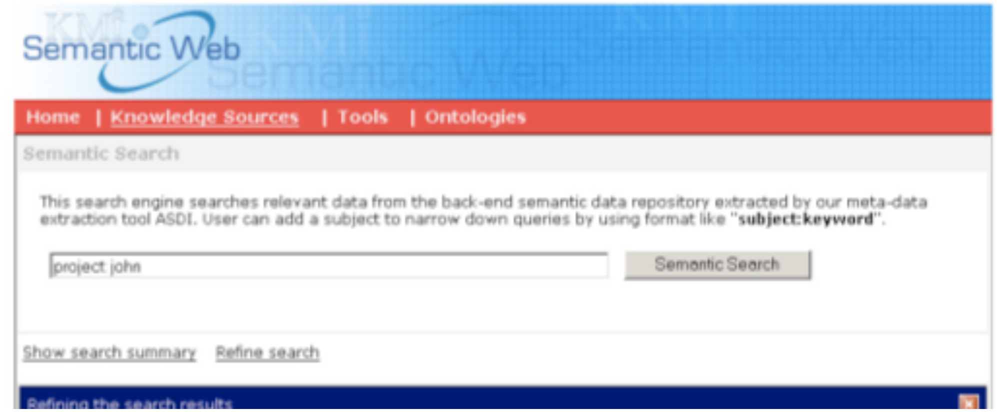
K-Search: retrieval



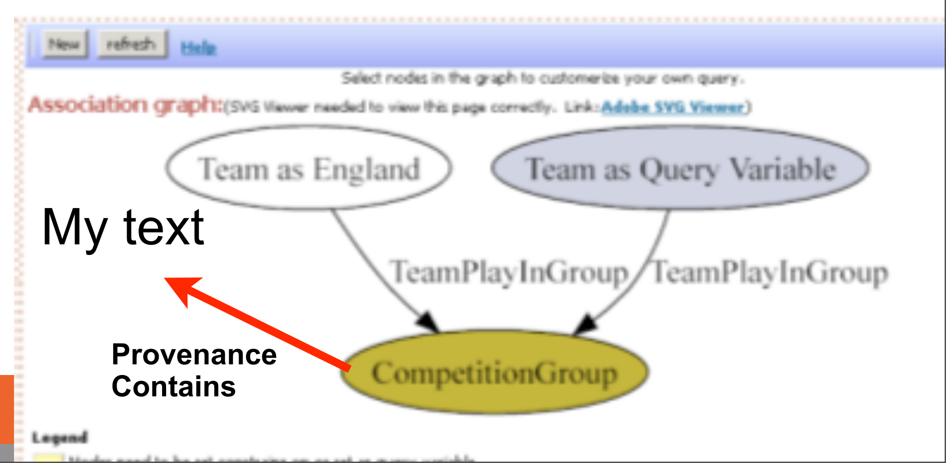
Hybrid Vs Pure Semantic Search

- It is possible to show that Hybrid search can be implemented in all the pure semantic approaches
  - With minimal change
  - Semantic search becomes a subclass of hybrid search

Implementing HS: What Search Strategy?

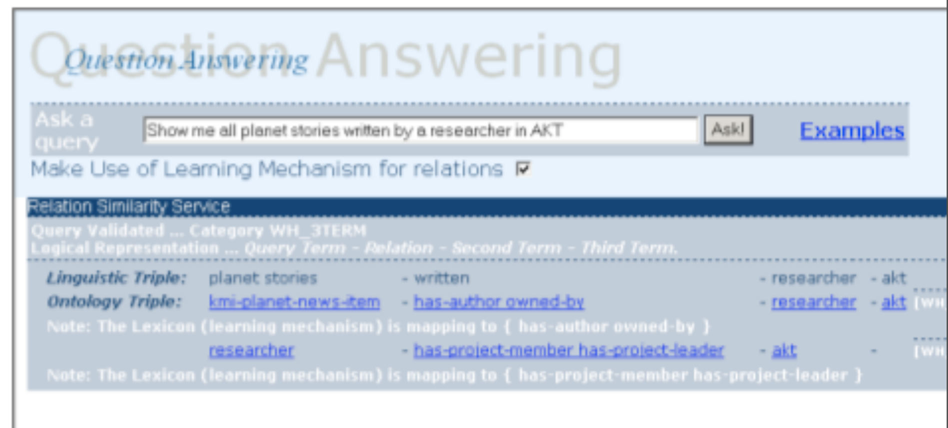


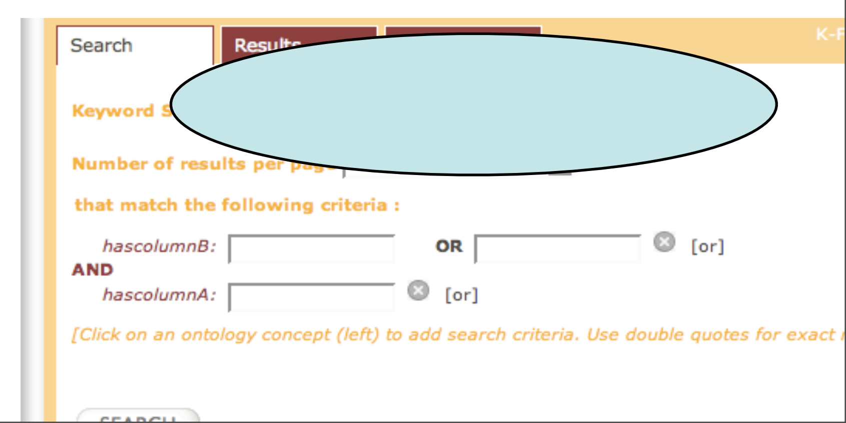
- Keyword-based approaches
  - Require translating all the keywords in order to perform the query
    - E.g. SemSearch
    - HS implemented by replacing keywords in the query with concepts in the ontology when possible while leaving the rest for pure keyword-based searching
    - Keywords in context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



Search Strategy (ctd)

- A natural language approach
  - E.g. Aqua
  - HS suported by recognising expressions like
    - "and the document contains…"
    - And its description contains
- Form-based approaches
  - HS supported by introducing
    - Keyword Search field
    - Enable keyword Matching on fields





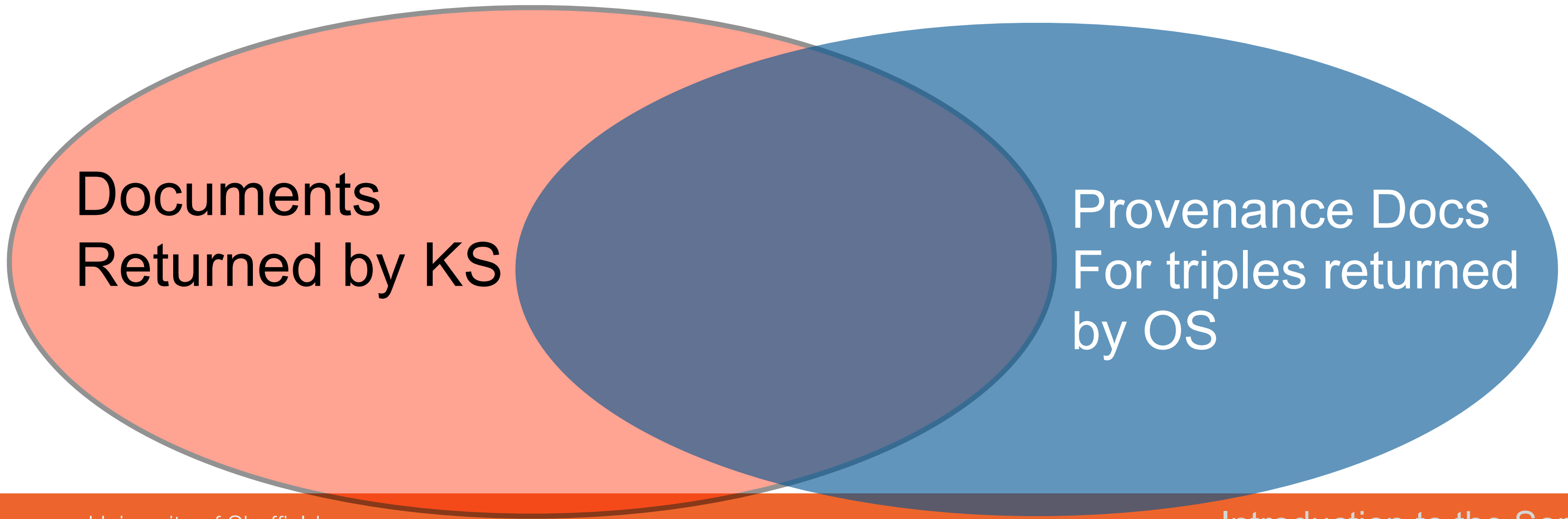
Result Merging

- 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

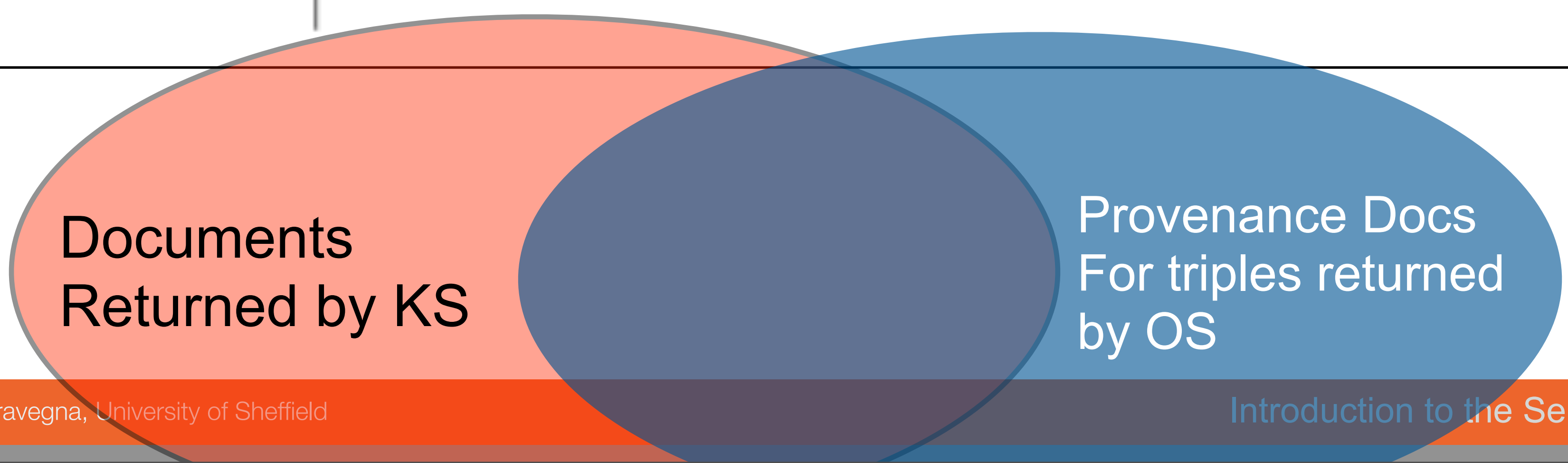
```
HybridSearchUriSet= KSDocUriSet ∩ OSDocUriSet
```



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

$$\texttt{HSTripleSet} = \left| \begin{array}{l} \text{All } \texttt{triples} \in \texttt{OSTripleSet} \\ \texttt{Where Provenance(triple}^{i}\texttt{)} \in \texttt{KSDocUriSet} \end{array} \right|$$



Ranking for Document retrieval

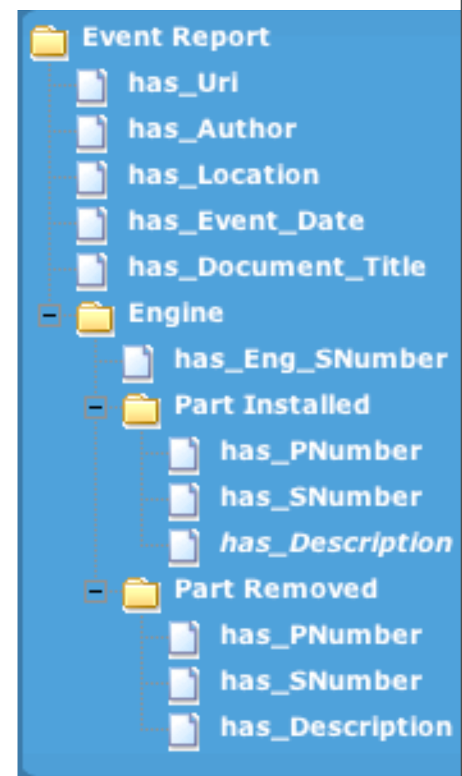
- Effective ranking is extremely important for a positive user experience
- Different ranking methods are possible
  - Document based
    - ability to match the keyword-based query
    - the keywords used in anchor links
    - the document popularity (given by link-based weights)
  - Knowledge Based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

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



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

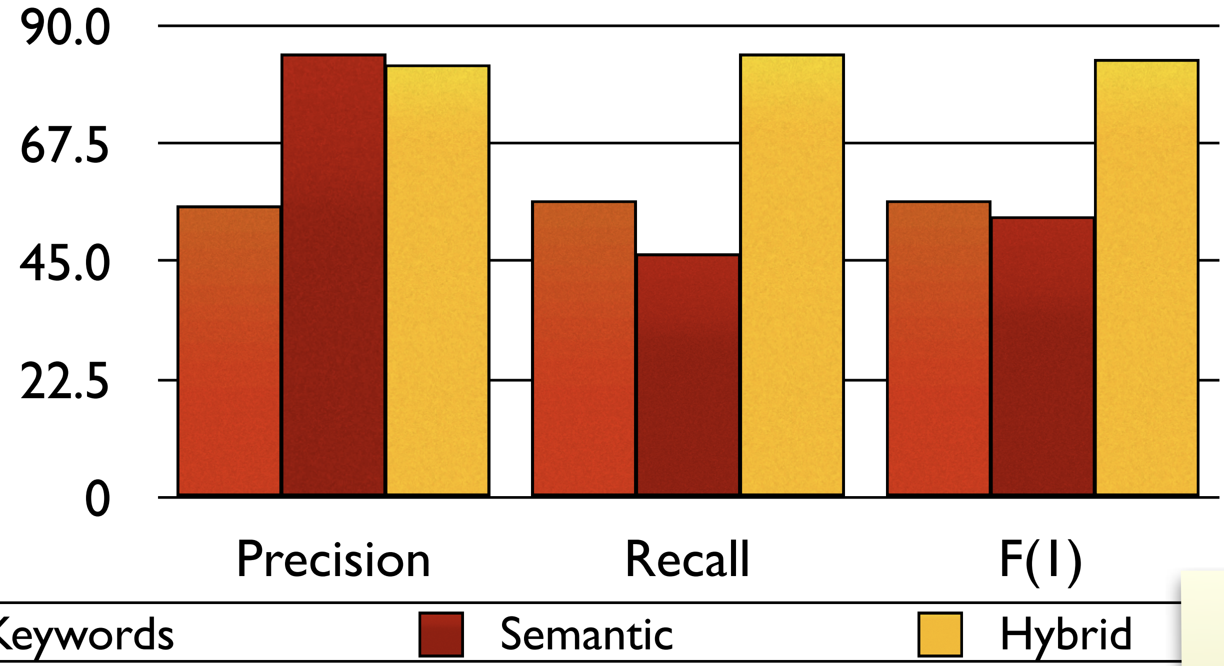
- Form-based implementation of hybrid search initially created for Jet Engine Designers
- It enables
  - Document querying
  - Knowledge querying
    - Including quantification of unstructured information

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

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

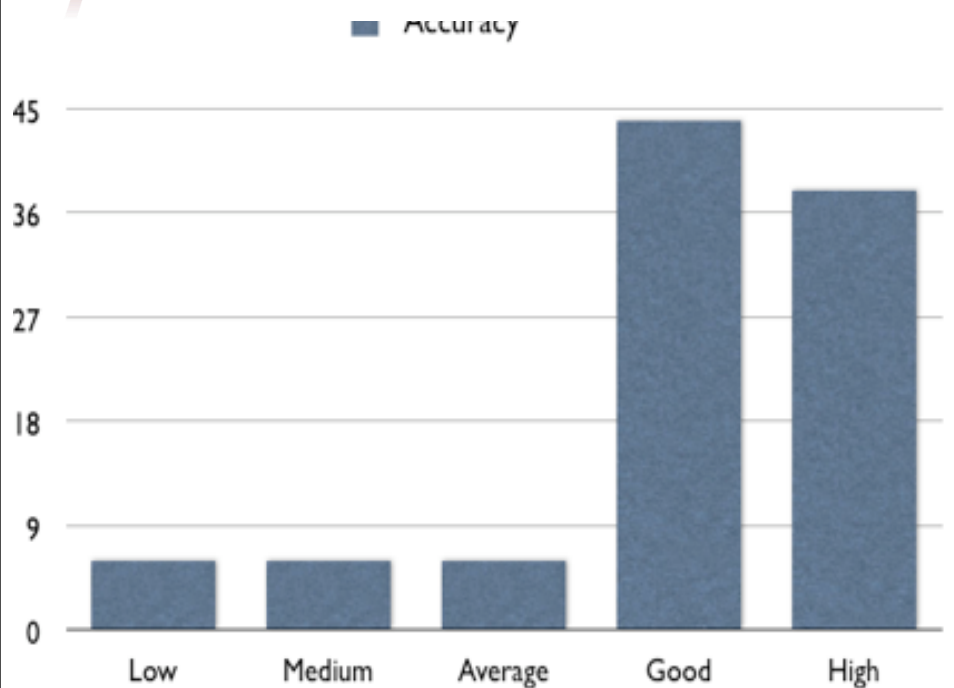
Final User Evaluation

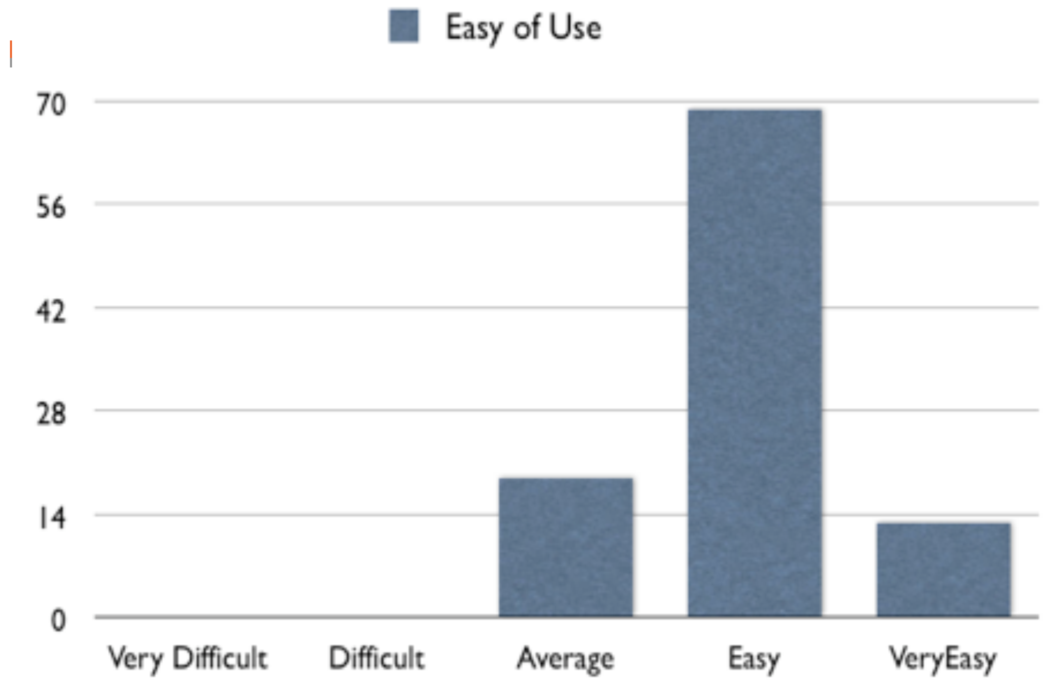
- Goal: verifying suitability for use in a real environment
  - 32 users Rolls-Royce engineers from different parts of the company
  - 90 minutes of test
    - Short introduction
    - 3 monitored tasks
      - One given (including solution)
      - One given (no solution)
      - One free task
  - Availability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

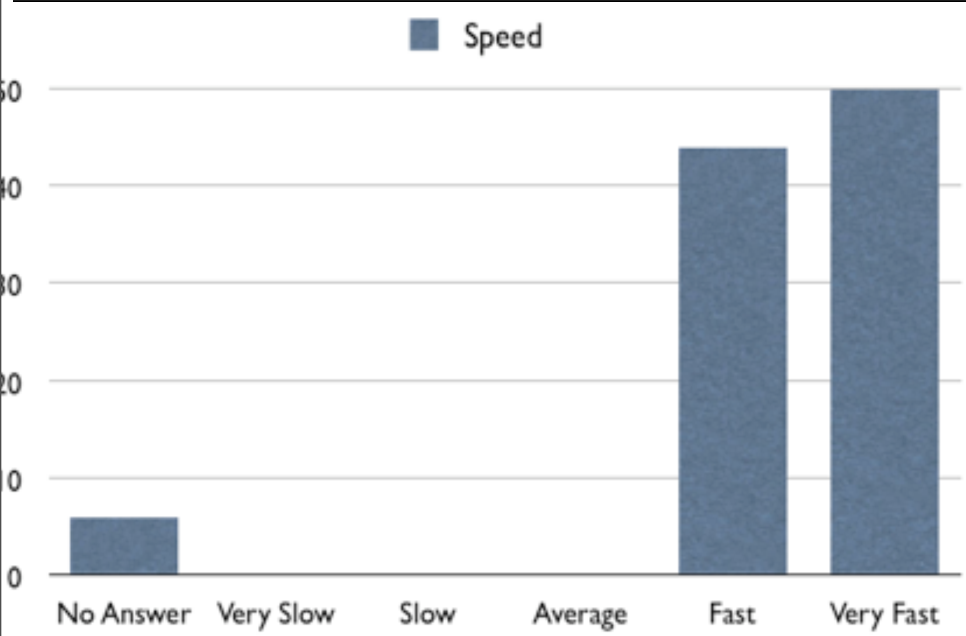
Evaluation Questions

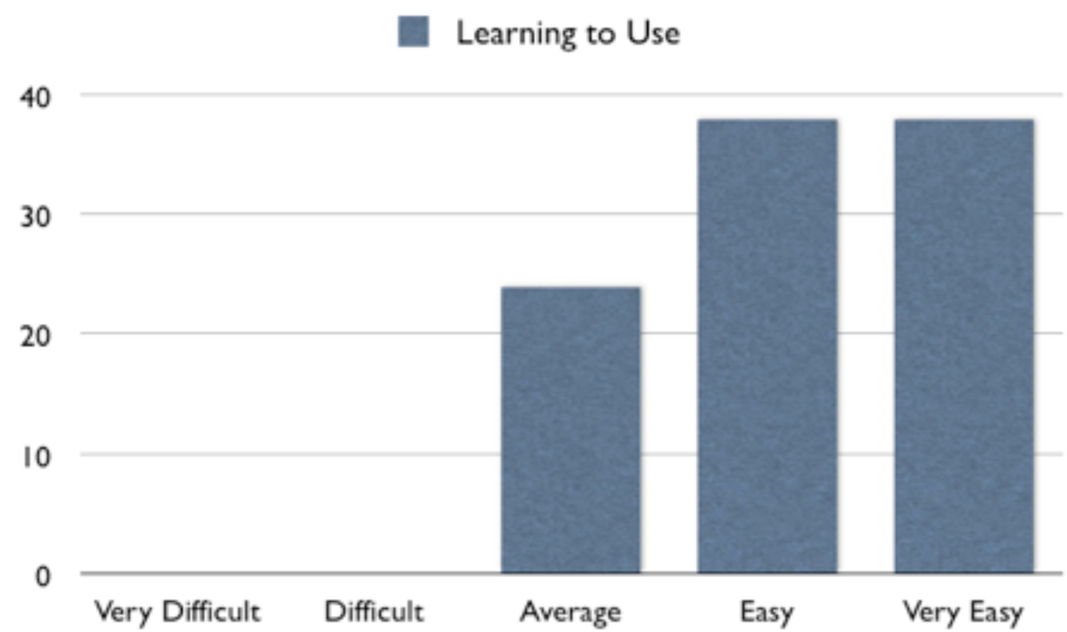
- Do user understand the hybrid paradigm?
- Are they able to search using HS?
- Do they actually use HS when confronted with a real searching task?
- Would the users be willing to use the system for their everyday work?

Liked by users?









Liked by Users?

- Finalist of Rolls-Royce Director's Creativity Award 2007
  - Voted by senior 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
      - Different paradigms for search

Future Work & Challenges

- Multidisciplinary research for annotation
  - 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)

- Modelling:
  - 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!

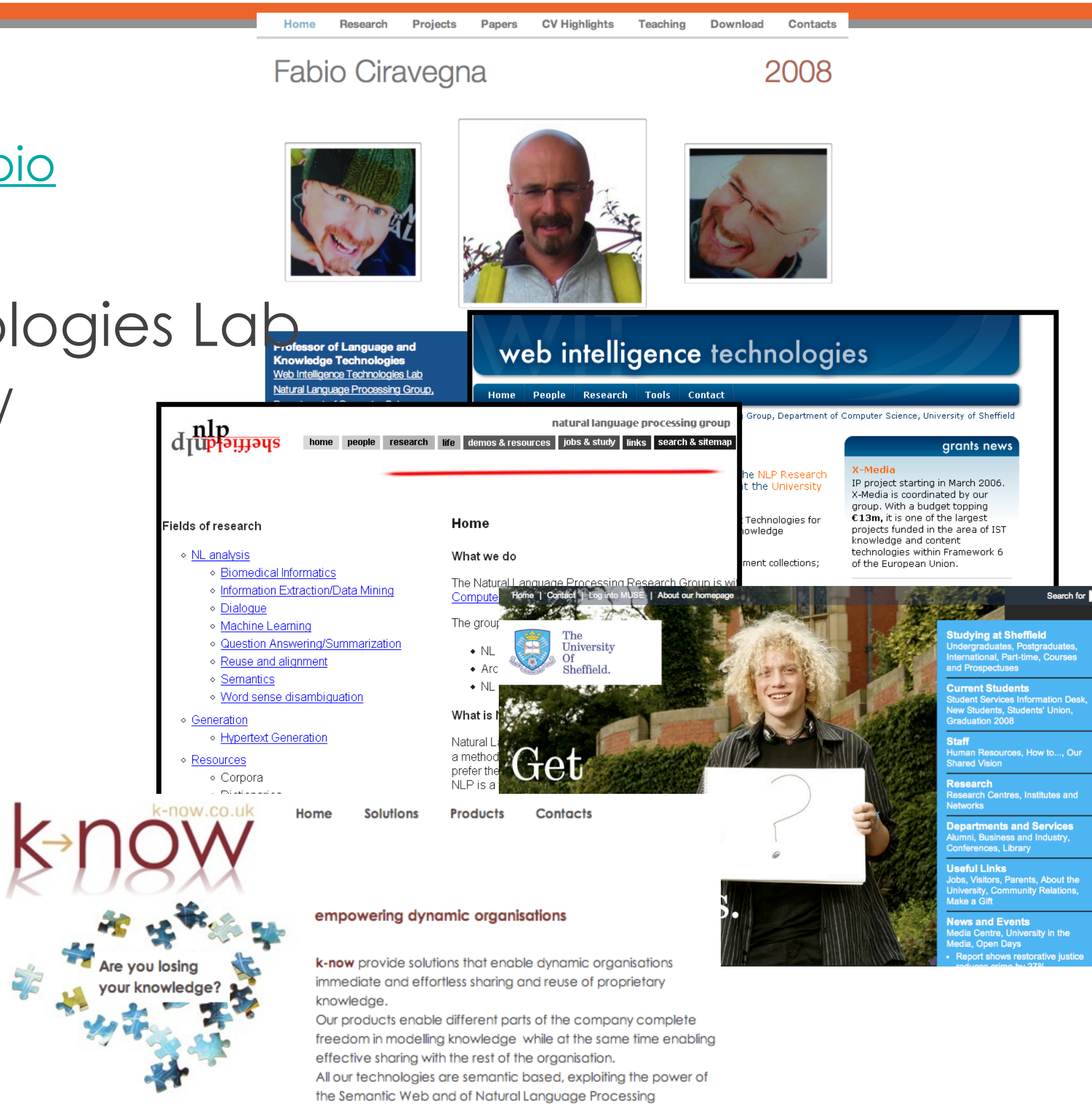
A final thought

- These technologies allow easy collection of 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?
      –

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
- K-Now Technologies
  - www.k-now.co.uk



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.