The University Of Sheffield.

ISWC 2007

Semantic Web Technologies for Knowledge Management in Large Distributed Organisations

Professor Fabio Ciravegna
Intelligent Web Technology Lab, Natural Language Processing Group
Department of Computer Science, University of Sheffield
http://www.dcs.shef.ac.uk/~fabio/

Sponsored by

www.3worldt.org

Sponsored by

X-MEDIA

www.x-media-project.org

The University Of Sheffield.

Copyright Notice

- These slides were presented during the International Semantic Web Conference (ISWC 2007) in Busan, Korea, November 2007 (http://www.iswc07.org/)
- They can be reused for personal and educational reasons only
- Updated versions and full conditions of use can be found at:
  - http://www.dcs.shef.ac.uk/~fabio/ISWC_tutorial/

ISWC 2007

The University Of Sheffield.

Outline of Tutorial

- 9.00-9.15 Introduction and scene setting
  - Issues in knowledge management in large organisations
- 9.15-9.45 Ontologies and ontology engineering
- 9.45-11.30 Semantic web technologies for knowledge acquisition
  - 10.30-11.00 Coffee break
- 11.30-12.15 Semantic web technologies for knowledge sharing, reuse and retrieval
- 12.15-12.30 Conclusion and future work
- 12.30- Discussion

ISWC 2007

Knowledge Workers' Challenges

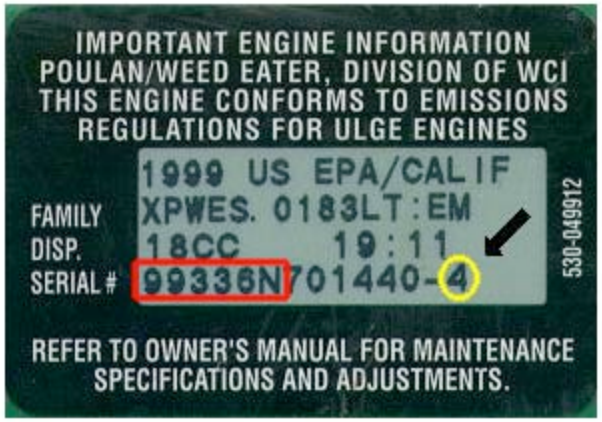
- Gathering knowledge relevant to a task or problem
  - it may be distributed across different storage systems and different media
- Analysing the knowledge they have gathered and make sense of it
- Sharing knowledge with their colleagues
- Keeping track of the process
  - by being aware of what one is doing, what one needs to do next, and what others are doing
- What to search for, what analysis is needed and who to share with
  - depend on the task in hand and the current stage of the process

Uren *et al.* 2007: X-Media Deliverable 4.1: Specification of Knowledge Sharing Systems

An Example of Knowledge Management

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

© Rolls-Royce plc



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

4 = Product Type

Jet engine example

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

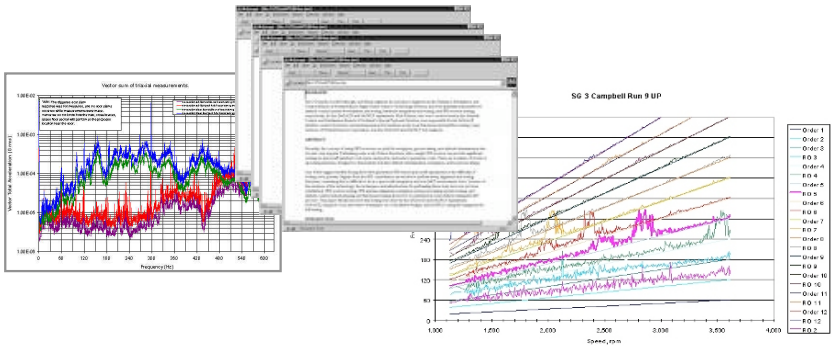
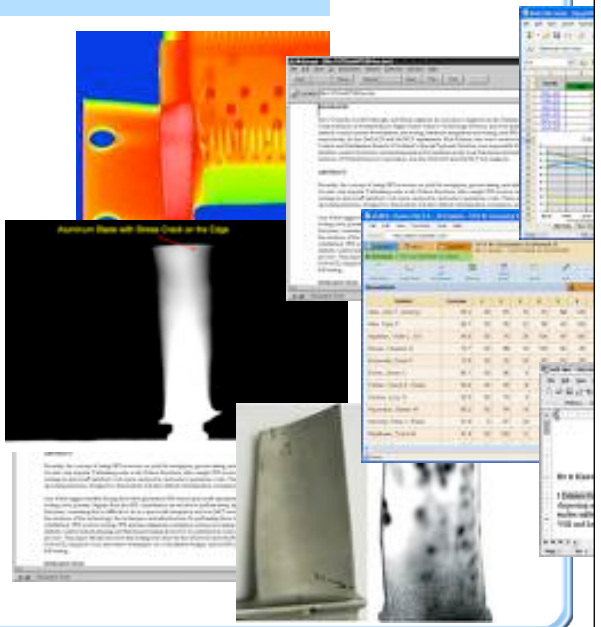


image © www.rolls-royce.com

Jet engine example (3)

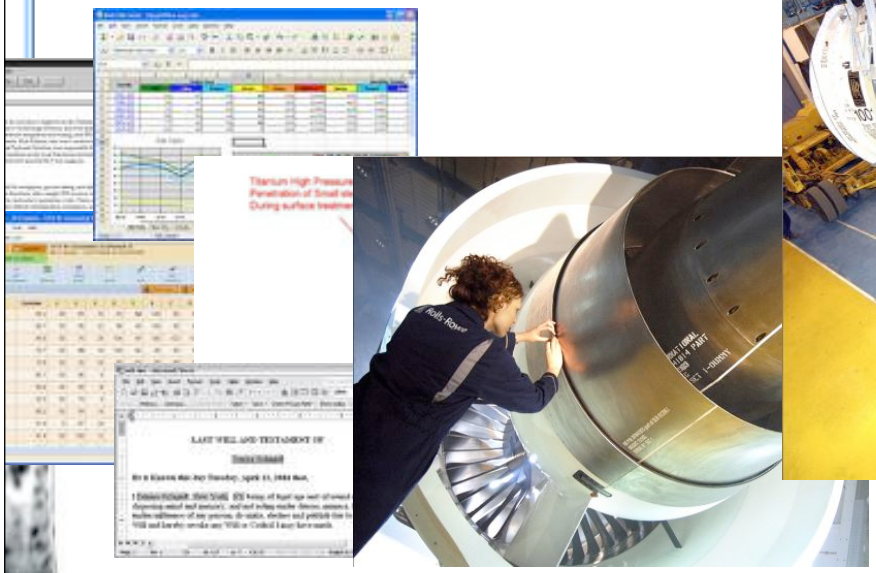
When engine is serviced (e.g. overhaul)

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



Jet engine example (4)

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

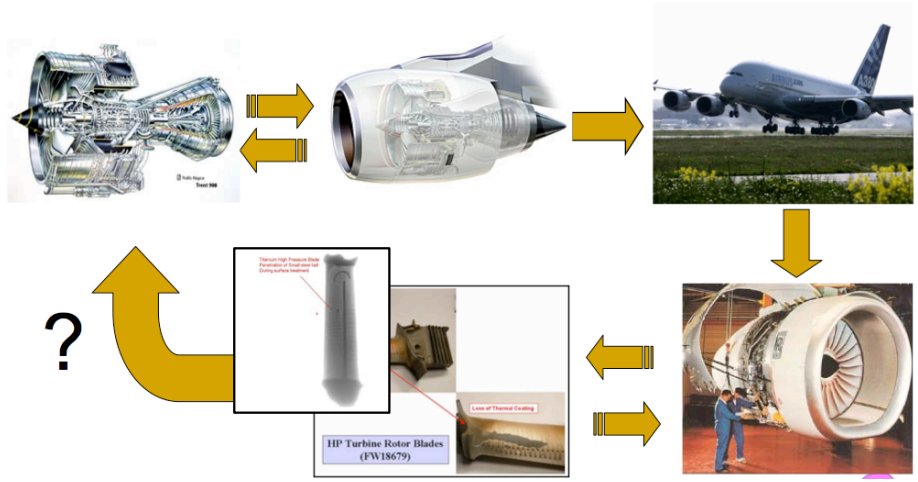


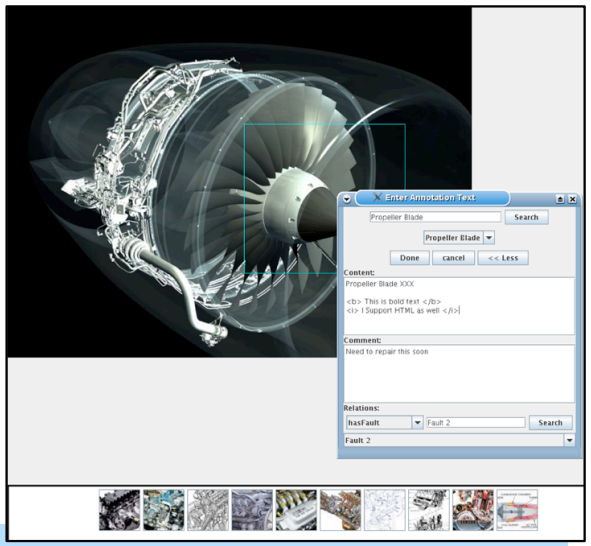
images © www.rolls-royce.com

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

Jet engine example (4)

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





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

What do we know and what we do not

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

Donald Rumsfeld

ISWC 2007

The University Of Sheffield.

Impact of Limited Sharing

- "The lack of efficient publishing capabilities for digital content costs organisations $750 billion annually due to wasted time spent by knowledge workers
  - seeking and capturing information necessary for them to do their jobs"

A.T. Kearney, Network Publishing Study, 2001
http://www.computeruser.com/news/01/04/17/news1.html

- According to S. Feldman:
  - 15%-35% knowledge worker time spent searching information
  - 50% of searches are successful (=50% fail)
  - 21% knowledge workers find information they need 85-100%

S. Feldman The high cost of not finding information. KMWorld Volume 13, Issue 3, March 2004.

ISWC 2007

Failing factors: Technical Issues

- Information scattered in multiple repositories
  - No one really knows which information is available and/or where
  - There isn't a single access point to information
  - Even a company-wide keyword searching facility is often inexistent
- 80-85% of a company's knowledge is unstructured
  - i.e. expressed in some forms of natural language or images/videos
- Information overload
  - Growing archives
  - Cost of storing very low
    - Video and 2D/3D image storing a reality

Failing Factors: Human Issues

- Everyone is a database designer
  - Everyone can create a database in some hours
    - Typically ill-designed
  - Some companies are Excel-based
    - Difficult searches
    - Archives do not scale to large size
      - time bomb (!)
- Everyone is a searcher! ...but with no training
  - Where to look, how to search, how to judge quality, when to stop…

The University Of Sheffield.

Management of What Type of Knowledge?

- Internal Knowledge
  - Need: capturing and sharing
  - e.g. How to design a product
- Focused external knowledge
  - Need: capturing, understanding, digesting, trusting and sharing
  - e.g. report of faults written by car garages
- External information
  - Need: capturing, understanding, contextualising, digesting, trusting and sharing
  - e.g. Information in Web pages
  - e.g. pictures provided by citizens in an emergency scenario

Issues: limitation of current KM systems

- Lack of Contextualisation for People, Processes and Technology
  - Current technologies tend to provide functionality in isolation from the processes and teams in which an individual knowledge worker plays a role
- Lack of Support for Cross Media and Cross Resource Sharing
  - Typically knowledge from a wide range of resources in different formats has to be brought together to solve a problem

Uren *et al.* 2007: X-Media Deliverable 4.1: Specification of Knowledge Sharing Systems

Limitation of current KM systems (ctd)

- Knowledge Generation Requires Initial Investment
  - Systems for producing rich metadata typically require a lot of user effort, for example by annotating documents to provide training data
- The False Assumption that Knowledge is Certain
  - Metadata is usually handled as if it were certain, ignoring the possibility that it may be incorrect, inaccurate or out of date
- Knowledge Gathering Lacks Expressivity and Contextualisation
  - Simple keyword based search engines do not support the needs of knowledge workers either in the sophistication of search formulation or in longer term and exploratory aspects of knowledge gathering

Uren *et al.* 2007: X-Media Deliverable 4.1: Specification of Knowledge Sharing Systems

The University Of Sheffield.

Issues with Keyword Matching

- Corporate archives are difficult to cope with because
  - Ranking cannot use document interlinking as search engines do
    - Risk of random order:
      - The % of Excite users who examined only one page of results per query in 2001: 50.5%
      - By 2001, more than 70 percent of Excite users looked at two pages or fewer
  - Documents can be very short and keyword matching has been proven not to work effectively on short documents
  - Vocabulary is reduced
    - Relevant terms tend to be very frequent
      - installed, engine, aircraft, removed, hazard, category, nrep, pse, blade, replaced, hkg, esn, csn are present in 50% of jet engine event report
  - Synonyms are not captured by keyword matching
    - Fuel Metering Unit, FMU, Metering Unit, fmu701mk5, S/N3332223

The University Of Sheffield.

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"
- Queries:
  - "what events caused during maintenance in 2003 were due to control units?"
- Translated into a set of queries given by all the possible combinations of:
  - "maintenance + 2003 + control + unit" (24 queries)

The University Of Sheffield.

Measures of Evaluation

- Can we trust a system answer?

$$Precision = \frac{COR}{min(ACT, maxNo)}$$

  - COR= correct answer by system
  - ACT= no results returned by system

- Are we finding all relevant documents?

$$Recall = \frac{COR}{EXP}$$

  - EXP= no of documents relevant to the query in archive

The University Of Sheffield.

Results for keyword matching

- 56% of documents in the first 20 hits are relevant
  - Precision=56%
- 57% of relevant documents are in the first 2 pages
  - Recall=57%

- Keyword matching implies
  - Reading a large amount of irrelevant documents
  - Risking missing documents
  - It is impossible to *count* the events

The University Of Sheffield.

Use of Web Technologies for KM

- Most resources are available over Intranets.
  - The Web is the medium for access to multiple archives in a seamless way
    - It enables remote access independently from geographic distribution
    - it provides a common protocol for communication
    - it provides an interaction modality familiar to users
  - However:
    - Scale: dozen to hundred million documents
    - Security: risk of information leaks and hackers' attacks limits access
    - User disorientation: multiple points of access

Knowledge Acquisition

Knowledge Reuse

Knowledge Modelling

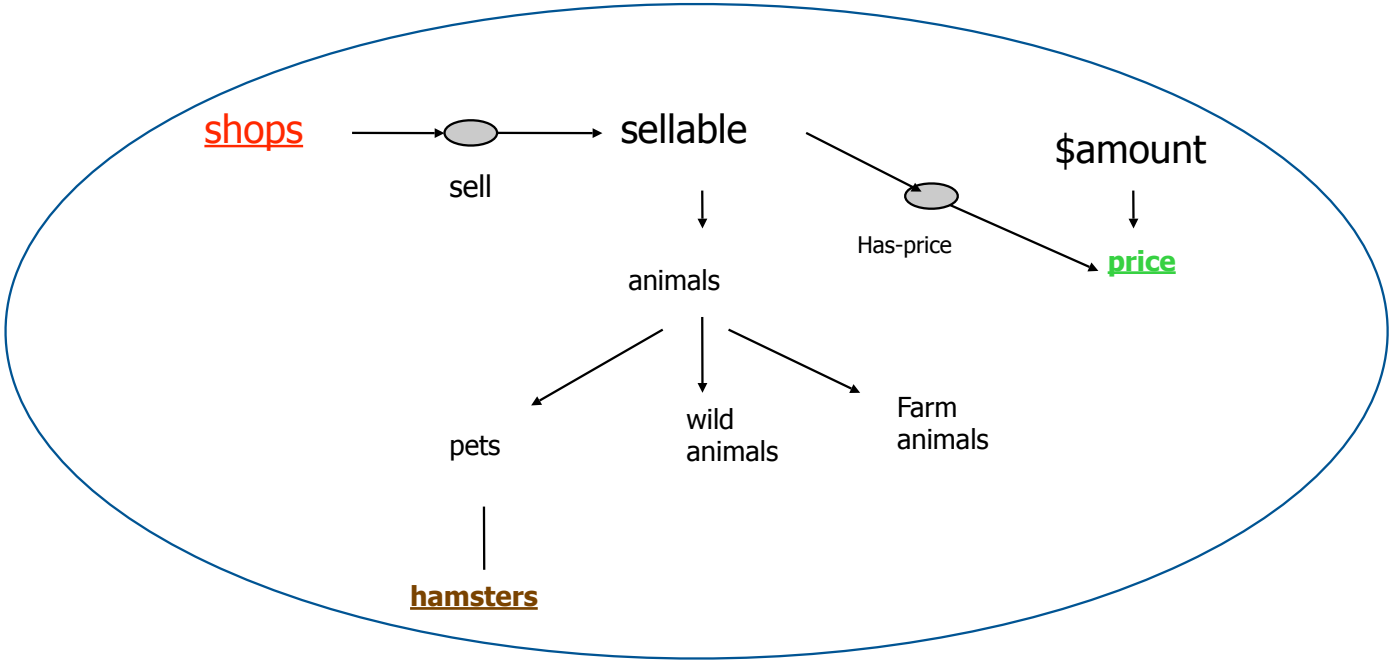
Knowledge Sharing

The Knowledge Life-cycle

The University Of Sheffield.

Steps in Knowledge Life-cycle we will see

- Knowledge Modelling
  - Ontology Engineering
    - Including also forms of: Acquisition + Reuse
- Knowledge Acquisition
  - Acquisition of information and knowledge from documents
  - Extracting and integrating information from existing archives
- Knowledge Sharing and Reuse
  - Enabling knowledge searching and process support



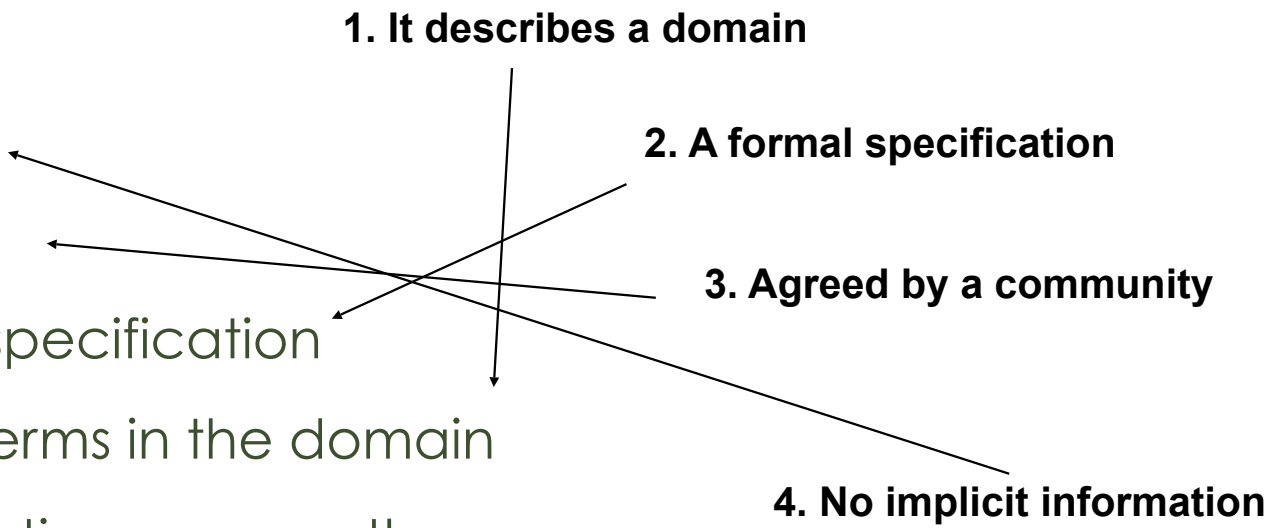
Ontologies and Knowledge Engineering

- Semantic web technologies: ontologies
  - use and role of ontologies: motivations and issues
  - cost of ontologies
  - scale of ontologies

The University Of Sheffield.

What is an ontology

- An
  - explicit
  - shared
  - formal specification
  - of the terms in the domain
  - and relations among them



Natalya F. Noy and Deborah L. McGuinness: Ontology Development 101: A Guide to Creating Your First Ontology
http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

ISWC 2007

The University Of Sheffield.

Ontology (2)

- An ontology defines a common vocabulary for agents (including people) who need to share information in a domain.
- It includes machine-interpretable definitions of basic concepts in the domain and relations among them.
- It is the main means of knowledge representation and interchange of information for the Semantic Web

The University Of Sheffield.

Why build an ontology?

- To share common understanding of the structure of information among people or software agents
  - E.g. for communication among sites in ecommerce
- To make domain assumptions explicit
  - Avoiding hardwiring into code or database schemas
  - Can be changed without changing code
- To enable reuse of domain knowledge
  - Including serendipitous use of knowledge

The University Of Sheffield.

Why? (ctd)

- To separate domain knowledge from the operational knowledge
  - Operational knowledge becomes more abstract
  - What works for cars will work also for trucks by just changing underlying ontology
- To analyse domain knowledge

The University Of Sheffield.

Ontology and Knowledge Base

- Elements in ontology
  - Classes or Concepts:
    - concepts in a domain of discourse
  - Slots (or roles or properties)
    - Properties of each concept describing various features and attributes of the concept
  - Facets (or role restrictions),
    - Restrictions on slots
- Knowledge base = instances
  - Instances or Individuals
    - Instances of concepts

The University Of Sheffield.

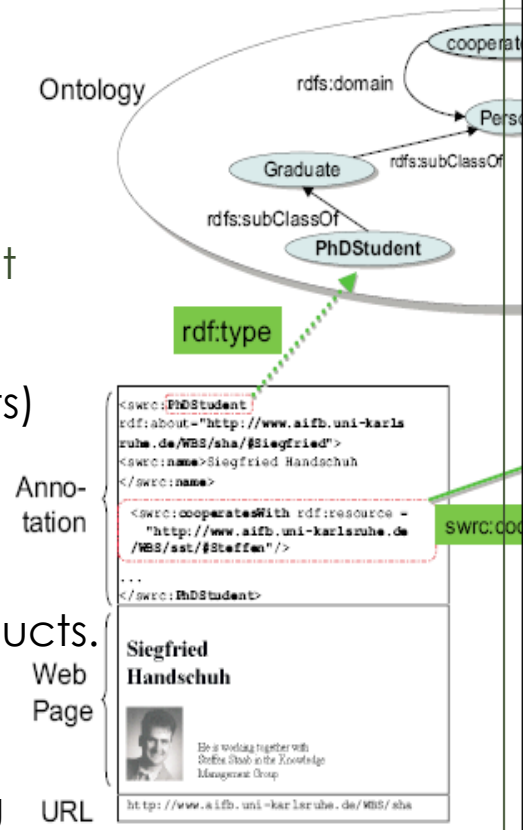
Ontology and Knowledge Base

- Ontology defines the domain in abstract terms
  - Types of objects, e.g. person and companies
- Knowledge base adds the specific individuals
  - Joe is-a person,
  - ACME ltd is-a company
- As an analogy think of
  - a database schema ~ ontology
  - actual content of database ~ instances

The University Of Sheffield.

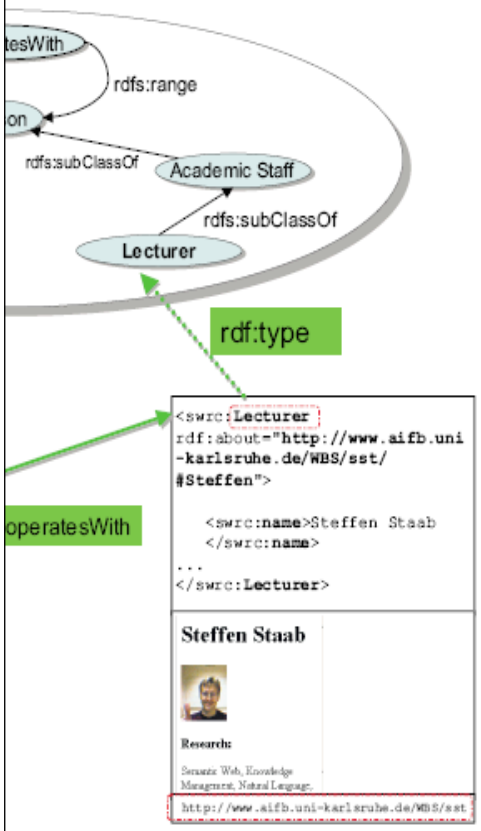
Ontologies and Knowledge Management

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



ISWC 2007

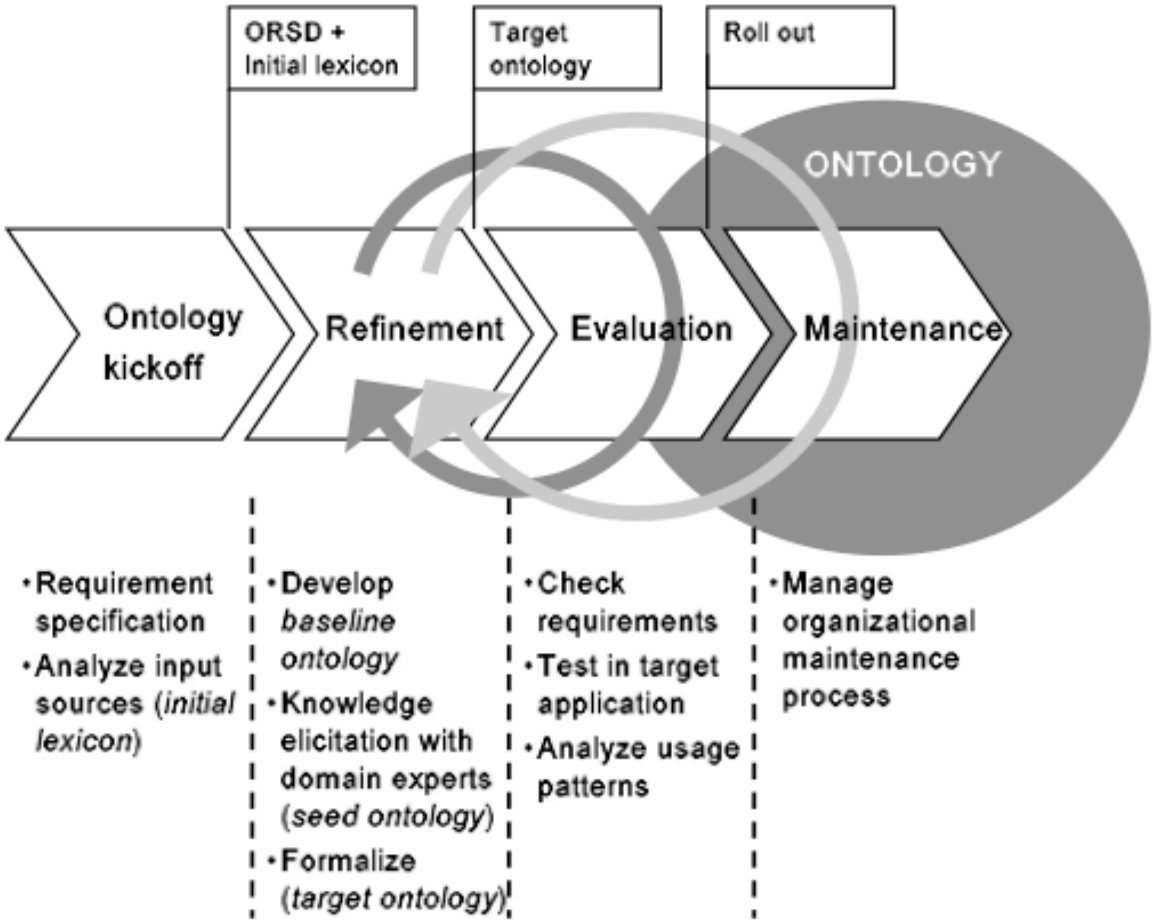
Ontologies & KM: issues



- Cost of knowledge engineering is very high
  - It requires commitment from company management
    - Chicken-egg problem
- Knowledge engineers are difficult to find
- Lack of engineering methodologies
  - What is the cost of an ontology?
- Cost of mapping information sources to ontology middle layer is high

The University Of Sheffield.

Ontology Engineering



Alexander Maedche, Steffen Staab, Nenad Stojanovic, Rudi Studer, York Sure: SEmantic portAL - The SEAL approach
In D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (eds.), Spinning the Semantic Web, pp. 317-359. MIT Press, Cambridge, MA., 2003.

The University Of Sheffield.

Steps in Ontology Engineering

- Requirements Analysis.
  - Domain experts and ontology engineers performs a deep analysis of the project setting w.r.t. a set of pre-defined requirements.
  - Includes:
    - re-use of existing ontological sources
    - extraction of domain information from text corpora, databases etc.
  - Result: ontology requirements specification document
    - Containing competency questions describing the domain and information about its use cases, the expected size, the information sources used, the process participants and the engineering methodology

Elena Paslaru Bontas, Christoph Tempich, York Sure : OntoCom: A Cost Estimation Model for Ontology Engineering
In: Proceedings of the 5th International Semantic Web Conference (ISWC 2006), November 5-9, 2006, Athens, GA, USA, LNCS. Springer.

The University Of Sheffield.

Steps in Ontology Engineering (ctd)

- Conceptualisation.
  - The application domain is modelled in terms of ontological primitives, e. g. concepts, relations, axioms.
- Implementation.
  - The conceptual model is implemented in a (formal) representation language, whose expressivity is appropriate for the richness of the conceptualisation.

The University Of Sheffield.

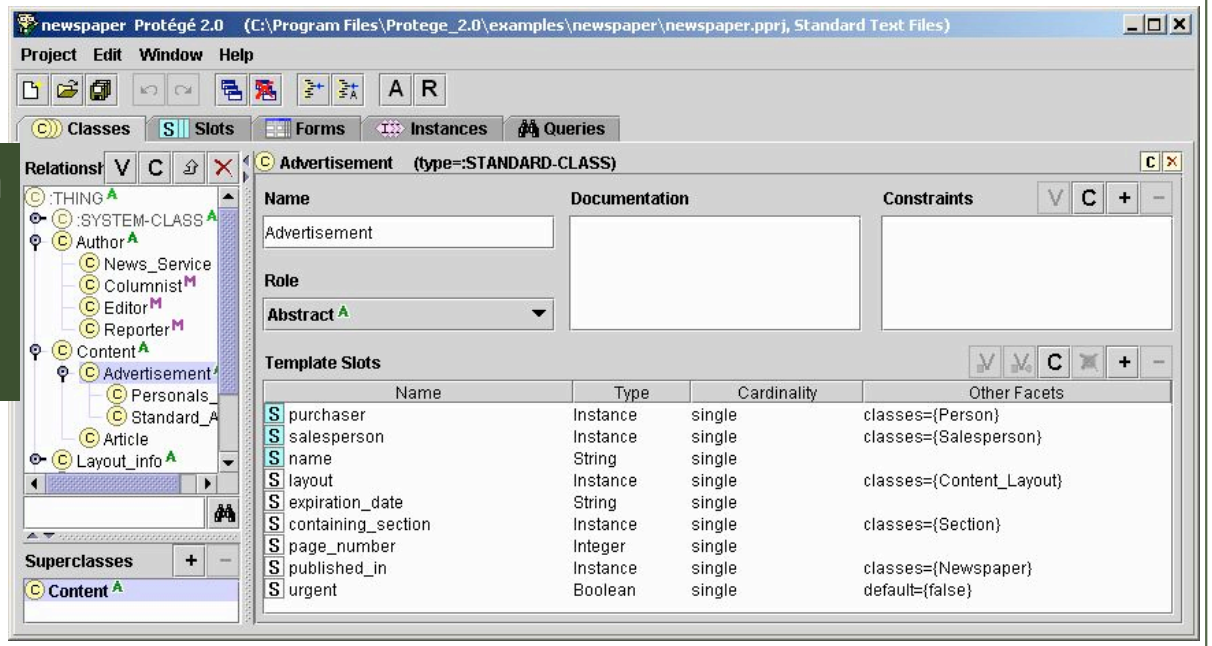
Steps in Ontology Engineering (ctd)

- Evaluation.
  - The ontology is evaluated against the set of competency questions.
  - The evaluation may be performed
    - automatically
      - if the competency questions are represented formally,
    - semi-automatically
      - using specific heuristics or human judgement
  - Result is a set of modifications/refinements at the requirements, conceptualisation or implementation level

Editing Ontology Tools: Protégé

- Protégé is a free, open source ontology editor
- Download at http://protege.stanford.edu/
- Protégé ontologies can be exported into a variety of formats including RDF(S), OWL, and XML Schema

Please note! Protégé is an editor! It does not support the whole knowledge engineering process



The University Of Sheffield.

Issues: Scale of Ontology

- Size of the ontology is a design issue for each application.
  - A jet engine has 30,000 parts;
    - Are they all to be represented as concepts?
      - What is a concept and what is an instance?
  - 20 different engine models.
  - Ontology + (knowledge base if some of them are instances)
    - large to very large if ontology contains representation of hyponymy and meronymy relations.
    - huge if we define precisely all the functional and positional relations among the 30,000 objects

Scale of KB up to billions of triples == size of ontology * number of engine types * number of events

ISWC 2007

The University Of Sheffield.

Size of Ontology (ctd)

- Large scale has important implications on the definition and management processes:
  - Careful hand-crafting in impossible
    - Large (possibly automatic) reuse of existing resources is necessary
  - Maintenance becomes complex
    - An ontology requires constant maintenance
    - Dedicated ontologists are uncommon in industry
      - Need of enabling users to update ontology

Ontologies in X-Media

- Ontology in X-Media is defined as a three layer structure:
  - Foundational ontologies (e.g. Dolce)
    - Main requirements: formal precision and generality, no maintenance required by team
  - Infrastructure ontologies
    - Communication between generic tools (like services, email, etc.), including multimedia ontology
      - e.g. an email has always a sender, a recipient and a subject/body
    - Requirements: formal precision, must be updated by expert ontologist
  - Domain ontologies
    - Describe the domain
    - Need not be so precise: they are often sloppy classification schemes (e.g. controlled vocabularies) or folksonomies or thesauri
    - Can be updated by trained experts in the domain

The University Of Sheffield.

Cost of Ontologies

- One of the major risks in ontology development is their cost
- Cost is split into:
  - PRODUCT-RELATED COST DRIVERS: account for the impact of the characteristics of the product to be engineered (i.e. the ontology) on the overall costs.
  - PERSONNEL-RELATED COST DRIVERS emphasize the role of team experience, ability and continuity w.r.t. the effort invested in the engineering process:
  - PROJECT- RELATED COST DRIVERS relate to overall characteristics of an ontology engineering process and their impact on the total costs

Elena Paslaru Bontas, Christoph Tempich, York Sure : OntoCom: A Cost Estimation Model for Ontology Engineering
In: Proceedings of the 5th International Semantic Web Conference (ISWC 2006), November 5-9, 2006, Athens, GA, USA, LNCS. Springer.

The University Of Sheffield.

Product Related Cost Drivers

- Domain Analysis Complexity
  - to account for those features of the application setting which influence the complexity of the engineering outcomes,
- Conceptualisation Complexity
  - to account for the impact of a complex conceptual model on the overall costs,
- Implementation Complexity
  - to take into consideration the additional efforts arisen from the usage of a specific implementation language

ISWC 2007

The University Of Sheffield.

Product costs (ctd)

- Instantiation Complexity
  - to capture the effects that the instance data requirements have on the overall process
- Required Reusability
  - to capture the additional effort associated with the development of a reusable ontology
- Evaluation Complexity
  - to account for the additional efforts eventually invested in generating test cases and evaluating test results
- Documentation Needs
  - to state for the additional costs caused by high documentation requirements

The University Of Sheffield.

Personnel-Related Cost Drivers

- Ontologist/Domain Expert Capability
  - to account for the perceived ability and efficiency of the single actors involved in the process (ontologist and domain expert) as well as their teamwork capabilities
- Ontologist/Domain Expert Experience
  - to measure the level of experience of the engineering team w.r.t. performing ontology engineering activities,
- Language/Tool Experience
  - to measure the level experience of the project team w.r.t. the representation language and the ontology management tools,
- Personnel Continuity
  - to mirror the frequency of the personnel changes in the team.

The University Of Sheffield.

Project-Related Cost Drivers

- Support tools for Ontology Engineering
  - to measure the effects of using ontology management tools in the engineering process
- Multisite Development
  - to mirror the usage of the communication support tools in a location-distributed team.

Towards the Web of Data

The University Of Sheffield.

Web of data and KM

- It is often said that the Semantic Web is rapidly evolving towards a Web of data
  - As opposed to Web of documents
- Web of documents
  - Feature: Web pages for human reading must be associated with machine readable data
  - Requirement: capturing information from unstructured documents
- Web of data
  - Feature: a large amount of information is contained in databases
    - e.g. eBay, Amazon.com, etc.
  - Requirement: mapping existing resources (e.f. database schemas) to an ontology and provide metadata together with data automatically
    - e.g. Semantic Amazon.com

The University Of Sheffield.

Reusing existing resources

- Reuse of existing sources
  - e.g existing ontologies/KB, Gazetteers or Database schemas or database content
- Existing resources:
  - Uncertainty
    - existing sources were generally invented for human reading
      - Are not 100% certain.
      - Trusting all background knowledge provided may decrease performance
  - Incompleteness
    - Background knowledge may only be available for part of the problem space.
  - Inconsistency
    - Pieces of knowledge from different sources can be conflicting

The University Of Sheffield.

(Re-)Using Different Ontologies

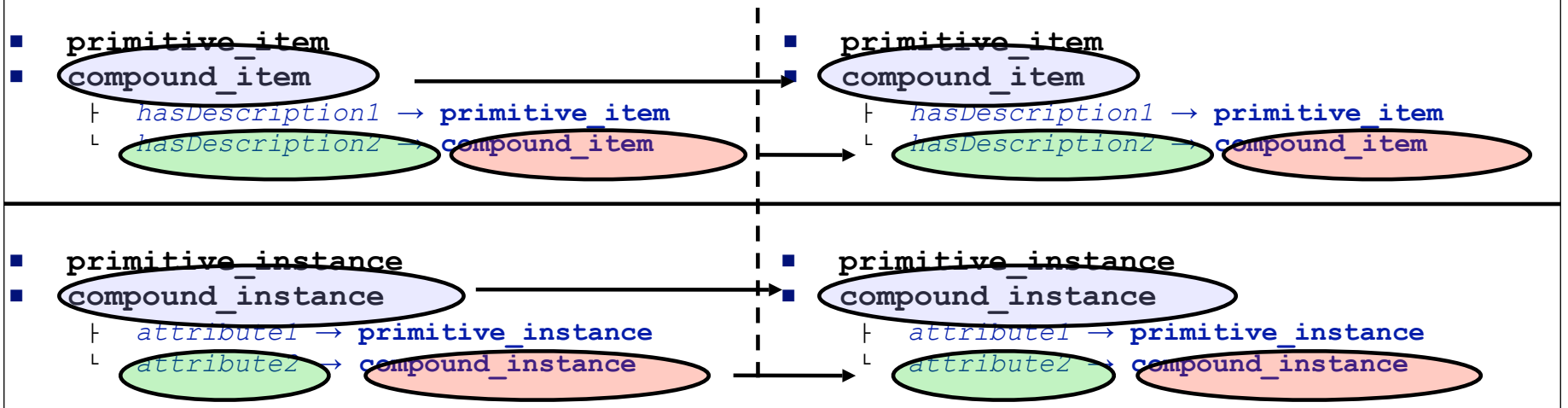
- Different communities use different domain representations
  - Design department uses designs and associated part lists
  - Service department uses illustrated part catalogue
    - The largely overlap but:
      - Use different URIs (part numbers)
      - Use different descriptions of the domain
        - In service what cannot be repaired is not described in details
        - In manufacturing what is outsourced is not described in details
- Ontologies as solution to map different resources (e.g. database schemas) to a common view

The University Of Sheffield.

Issues in mapping schemas

- The main difference between an db schema and an ontology
  - Ontology makes assumptions explicit
  - Database schema leave assumptions implicit
- Issue:
  - Mapping requires:
    - Making assumptions explicit (turn schema into ontology)
    - Mapping the two ontologies
      - Three main approaches
        - Top down (next slides)
        - Bottom up (starting from instances - see slides on integration)
        - Mixed

Ontology Mapping



- class to class mapping (*classMapping)*
- attribute to attribute mapping (*attributeMapping*)

A Slide by Adrian Mocan

http://www.inrialpes.fr/exmo/people/zimmer/SDK-meeting/Presentations/Adrian%20Mocan%20-%20WSMX%20Data%20Mediation.ppt

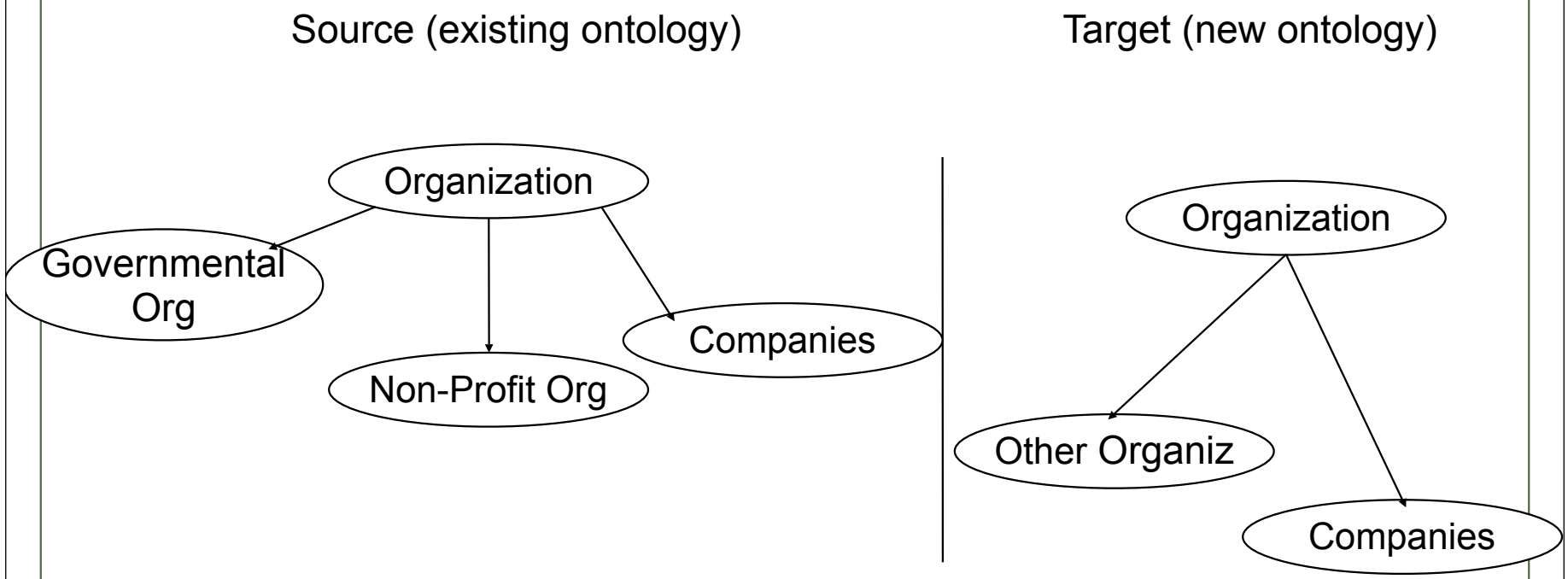
Ontology Mapping: Issues

- Difficulties in mapping concepts and properties
  - Non overlapping
    - Concept/Property X in DB Schema does not exist in Domain Ontology
      - Solution: extension of Domain Ontology to include it
    - Concept/Property X in Domain Ontology does not exist in DB Schema
      - Solution: none; missing information
  - Partially Overlapping
    - Concept/Property exists but with a slightly different definition
    - Next slide

Partially Overlapping Objects (1)

Specific to generic

- Easy case
  - Collapse all instances and subtypes of Governmental_Org and Non-profit_Org into Other_Organiz

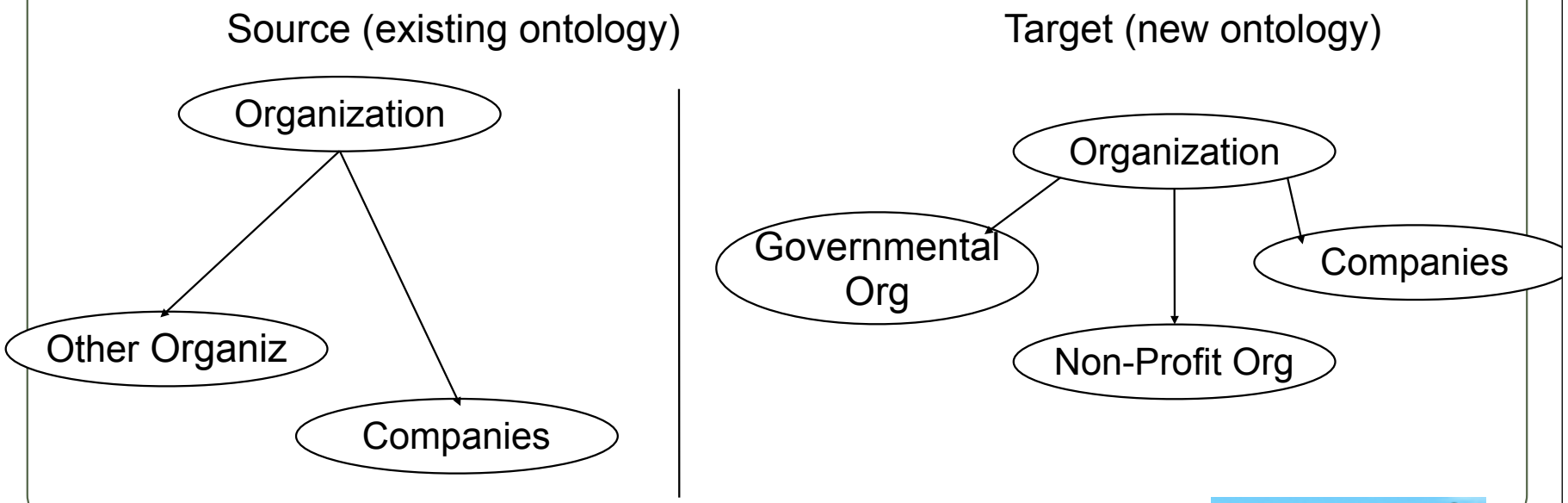


ISWC 2007

Partially Overlapping Objects (2)

Generic to Specific

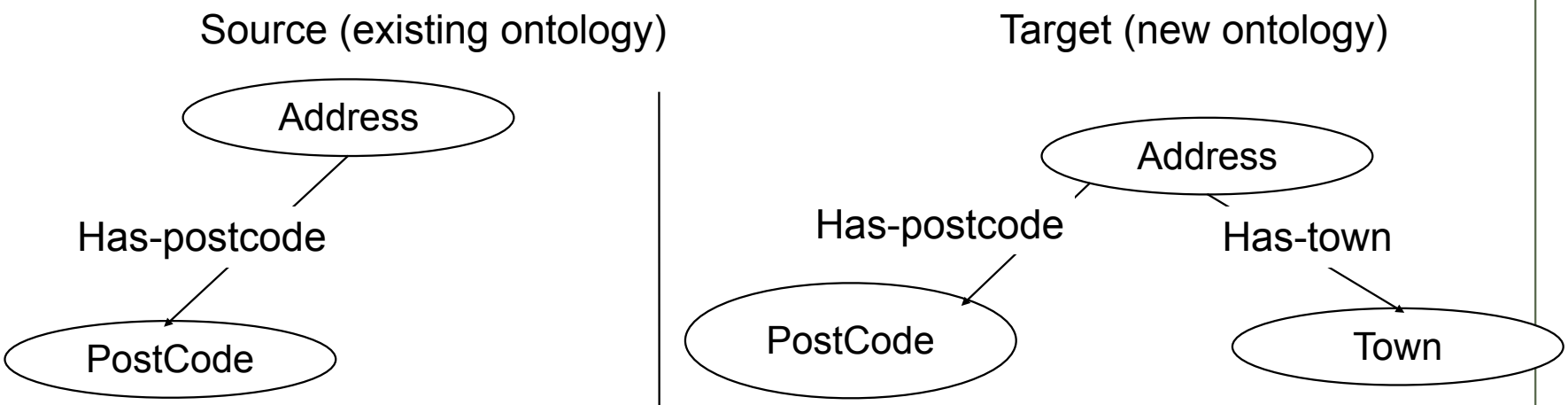
- Difficult case
  - How do we divide the Other_Organiz into Governmental_Org and Non-profit_Org into?
  - Manual mapping of instances/subtypes or modification of ontology

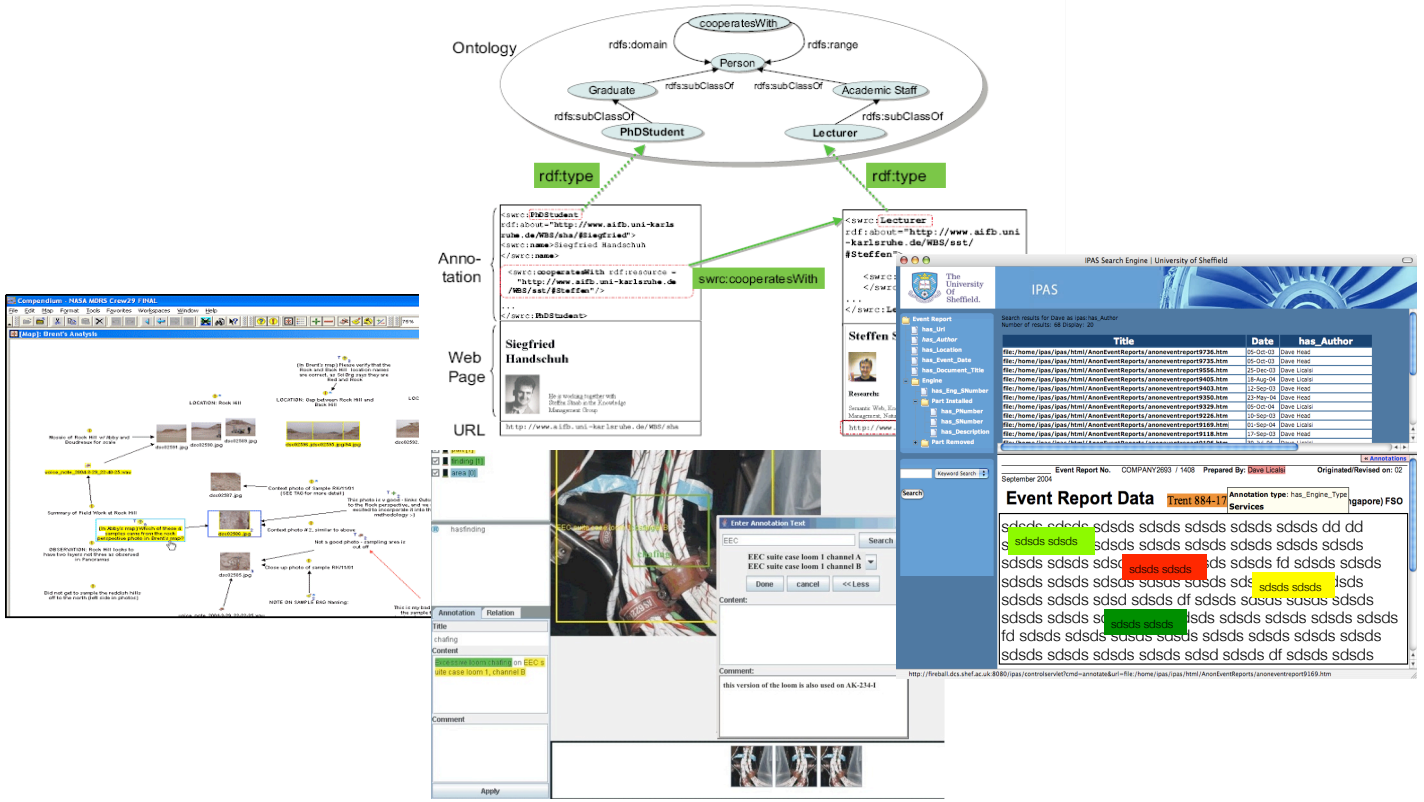


Partially Overlapping Objects (3)

A more complex case

- Difficult case
  - In the DB the address implicitly includes the town
    - E.g. substring
  - In the Domain Ontology the town is explicitly mentioned
  - No easy way to map



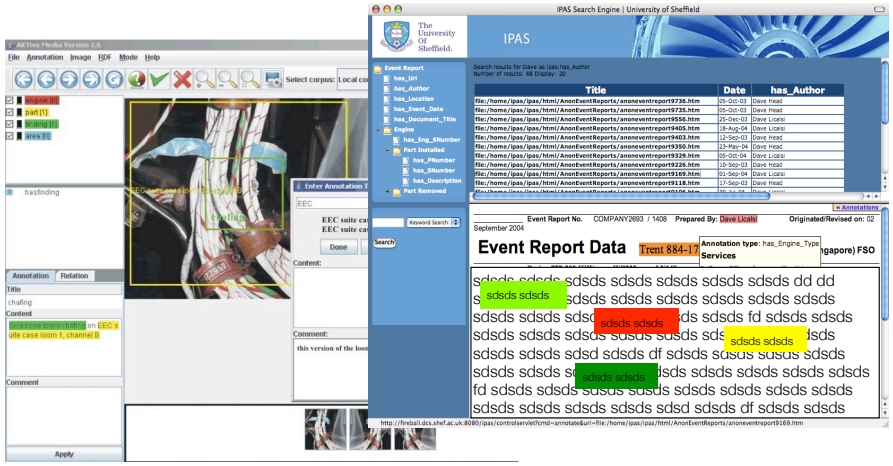


Requirements for Knowledge Acquisition

- issues in knowledge acquisition:
  - acquiring: what and what for?

Knowledge Acquisition

- 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



Requirements for KA: Web Size

- August 2005: Yahoo claimed to cover 20 billion pages
  - 19.2 billion web documents,
  - 1.6 billion images,
  - 50 million audio and video files.

http://google.weblogsinc.com/2005/08/08/yahoo-upgrades-indexes-claims-20-billion-objects/

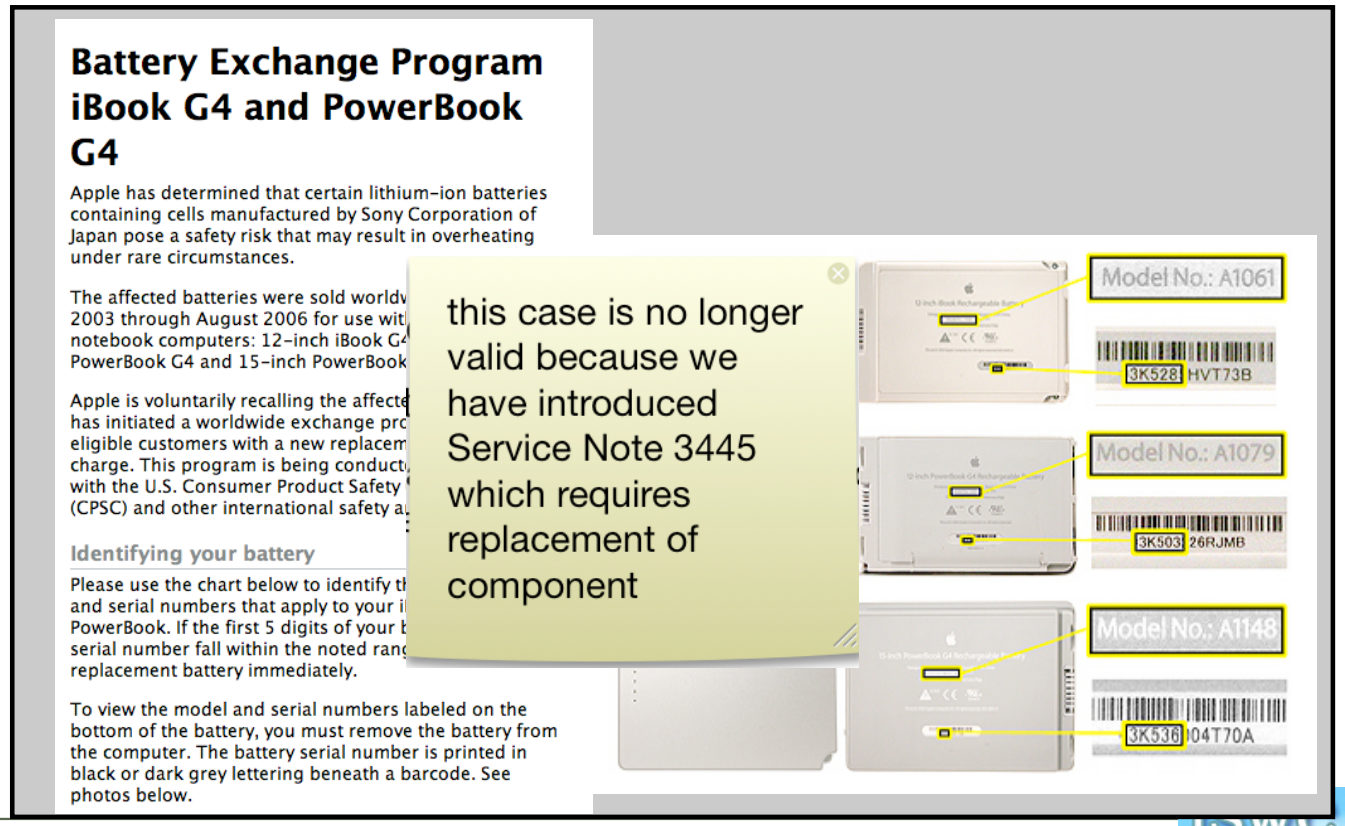
- Almost 375 petabytes (or 787.5 billion photographs) are produced each year
  - almost 2 times all printed material
  - yearly growth rate of 5%
    - Highest growth rate among different data types

Brilakis, I.: Content based integration of construction site images in aec/fm model based systems, PhD thesis, University of Illinois at Urbana-Champaign, 2005

Large intranets are following the Web's trends. Going towards hundreds of millions documents (web at end of the 90s)

Requirements for KA: Cross media

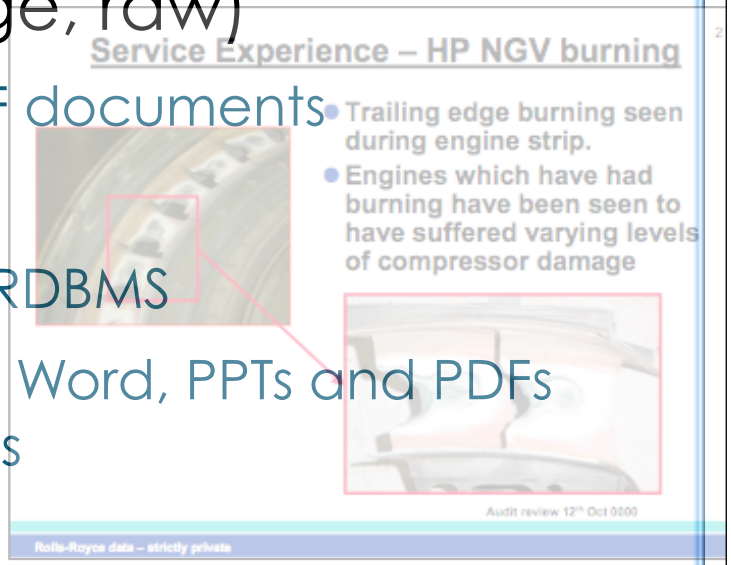
- Evidence is often distributed in different media;
- Knowledge in one medium does not carry the full evidence



Compound Documents & CM

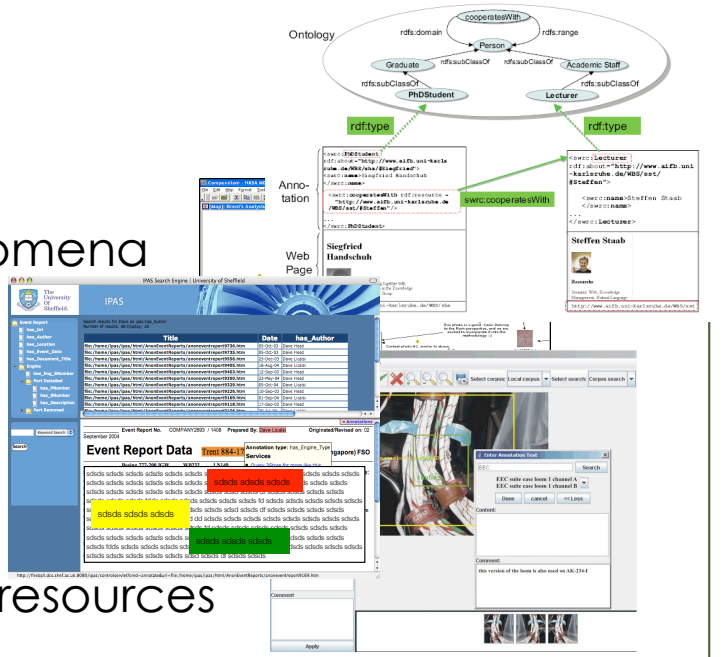
From Deliverable D8.2

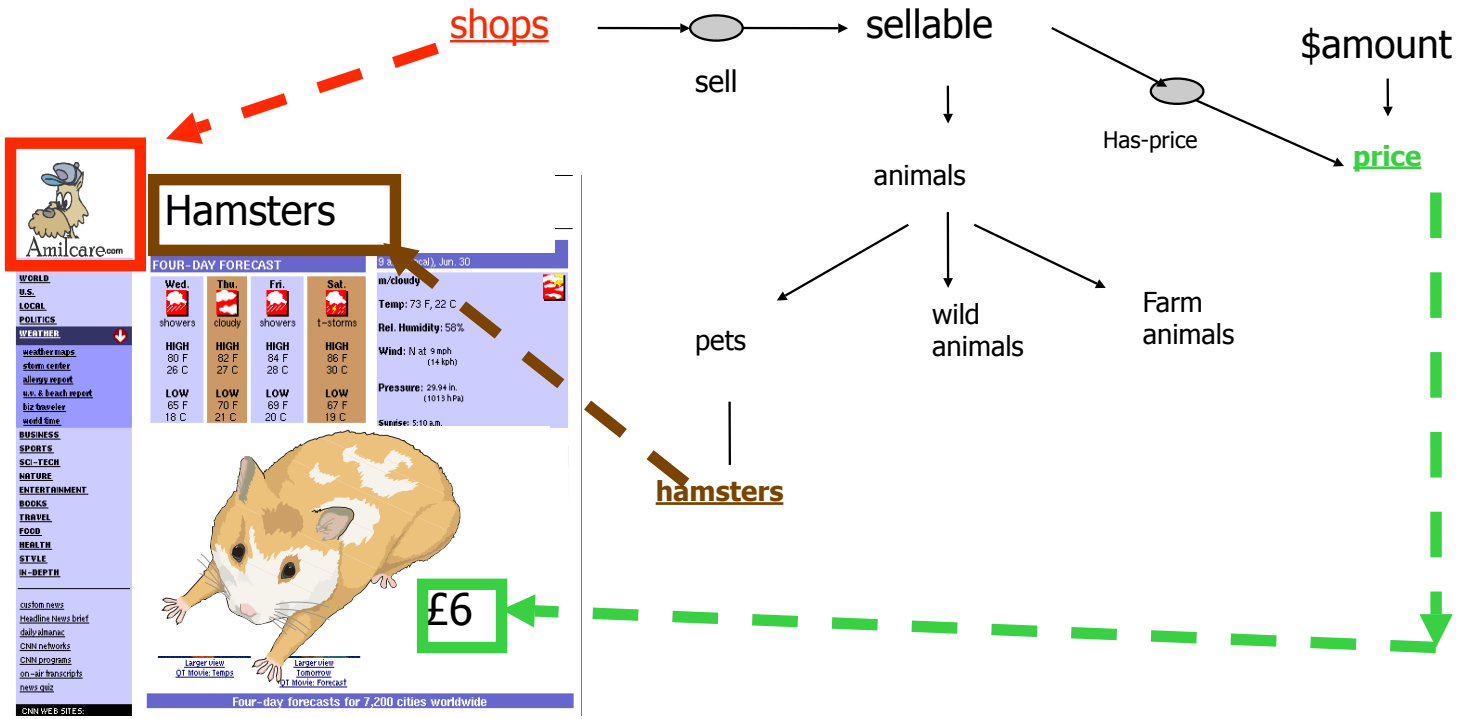
- Typical data objects (text, image, raw)
  - Text formats: Word, Excel, PPT and PDF documents
  - Images: Jpeg and Gif
  - Raw data: Measurements stored in a RDBMS
  - Cross-media: Compound documents: Word, PPTs and PDFs containing both text and Jpeg images
    - Portions semantically related to each 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



Requirements for SW: Robustness

- Required robustness in:
  - Knowledge representation
    - e.g. uncertainty and dynamic phenomena modelling
  - Against unexpected situations:
    - Coping gracefully with downtime of resources
      - What if a document disappears/server is down? (reasoning)
    - Preventing that a crash of an individual components leads to a whole system down.
    - Dealing intelligently with error propagation through the cascade of processors





SW for Knowledge Acquisition

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

ISWC 2007

The University Of Sheffield.

Semantic Web for Knowledge Acquisition

- Aims:
  - To acquire knowledge within and across media in a rich, semantically-oriented way
  - Outcome of acquisition 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

The University Of Sheffield.

Making Content Available

- 3 main methods of making the content available:
  - Ontology-based annotations
  - Free text annotations - Braindumps
  - Document enrichment

Vitaveska Lanfranchi, Fabio Ciravegna and Daniela Petrelli: Semantic Web-based Document: Editing and Browsing in AktiveDoc, 2nd European Semantic Web Conference, Crete, June 2005

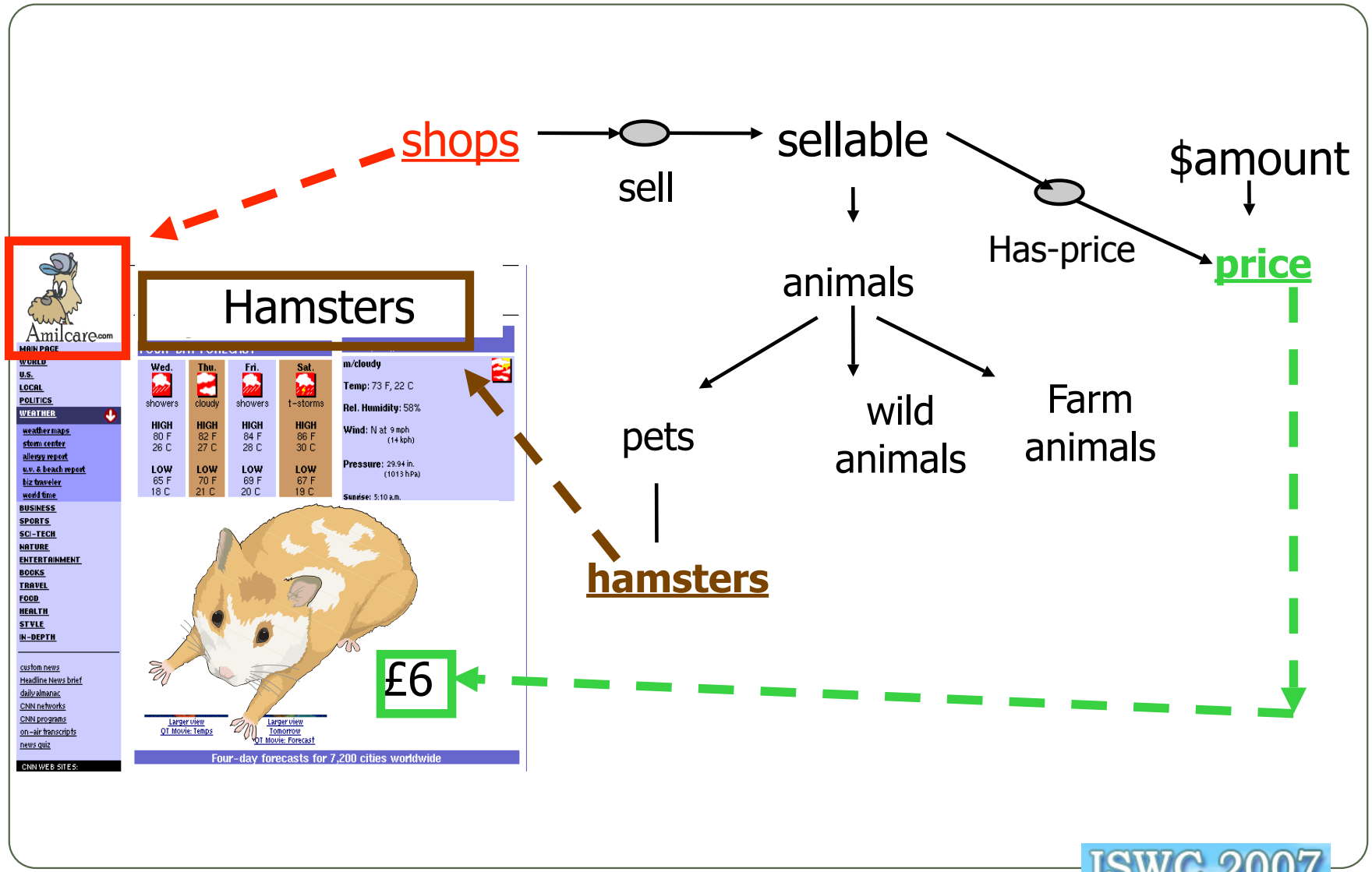
ISWC 2007

The University Of Sheffield.

Ontology-based annotations

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

Ontology-based Annotation



The University Of Sheffield.

Free text annotations - braindump

- Adding knowledge to documents
  - Via free – text comments (as in Word)
  - This is called braindump
    - The final document is just the final solution
    - E.g. the project for a new Jet Engine
      - During the discussion the working group will consider many alternative solutions
      - Those not selected are not in the final project
      - When next jet engine is designed, the group needs to know
        - What solutions were tried (use of titanium)
        - Why they were not adopted (e.g. too high a cost)
        - If the analysis is still true (titanium cost has decreased)
    - Adding further comments and associate Information
      - Annotea (Barstow et al 2001)
      - Semantik (Gilardoni et al 2004)

Braindump in a Legal Scenario

The University Of Sheffield.

Plain English in the Twenty Types of Legal Documents

By George H. Hathaway

To eliminate legalese from legal documents, you must do more than just criticize legal writing in general. You must zero in o
organized way, rather than randomly selecting a sample of legal writing and critiquing it. Therefore, we have grouped all legal
Figure 1.

We have published many articles about these 20 types of documents in our Plain English theme issues of November 1983 a
present. Now we will discuss: 1) the level of clarity at which each type of document is presently written; 2) improvements, if
are still unresolved...and why.

Resolutions1

Whereas, it is a privilege to congratulate Kyle David Gibbs [redacted] the rank of Eagle Scout...

Resolutions are written by the Michigan House and Senate to express a position on an issue. The resolutions are published b
simple, and therefore the main body of the resolution is usually clear. [redacted] solutions have always contained one persis

Trying to eliminate this word from resolutions illustrates the intractability of those who write legalese on purpose. So far the H
make up only 15% of the Legislature. This means six out of [redacted] ators are not lawyers. Furthermore, the clerks of the
doesn't the Michigan Legislature eliminate Whereas from its resolutions? There are still too many people (lawyers, non-lawye
symbol of power and prestige.

Statutes2

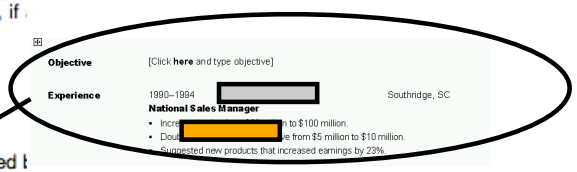
Michigan statutes are written by the Legal Division of the Legislative Service Bureau. (Of course, the drafters often do not ha
the Michigan Legislative Service pamphlets and each year in the Public and Local Acts of Michigan. In 1994 we reviewed ne
Service Bureau for its work.

Executive Orders3

Whereas, Article V, Section 2, of the Constitution of the State of Michigan of 1963 empowers the Governor to make changes
units which he considers necessary for efficient administration; and....

These orders are written by the Governor's legal counsel and are published each month in the Michigan Register. They are di
format for executive orders has not changed in the last 100 years. It has always contained much legalese. We are told that it
orders carried as much weight as legislative statutes, administrative rules, or case opinions. They believe that if the orders a
have a better chance of being followed. But this is what critics of legal writing have always chargedNthat lawyers write legale

Why we used these references



Why we DID NOT use other references

quinary

DOT.KOM

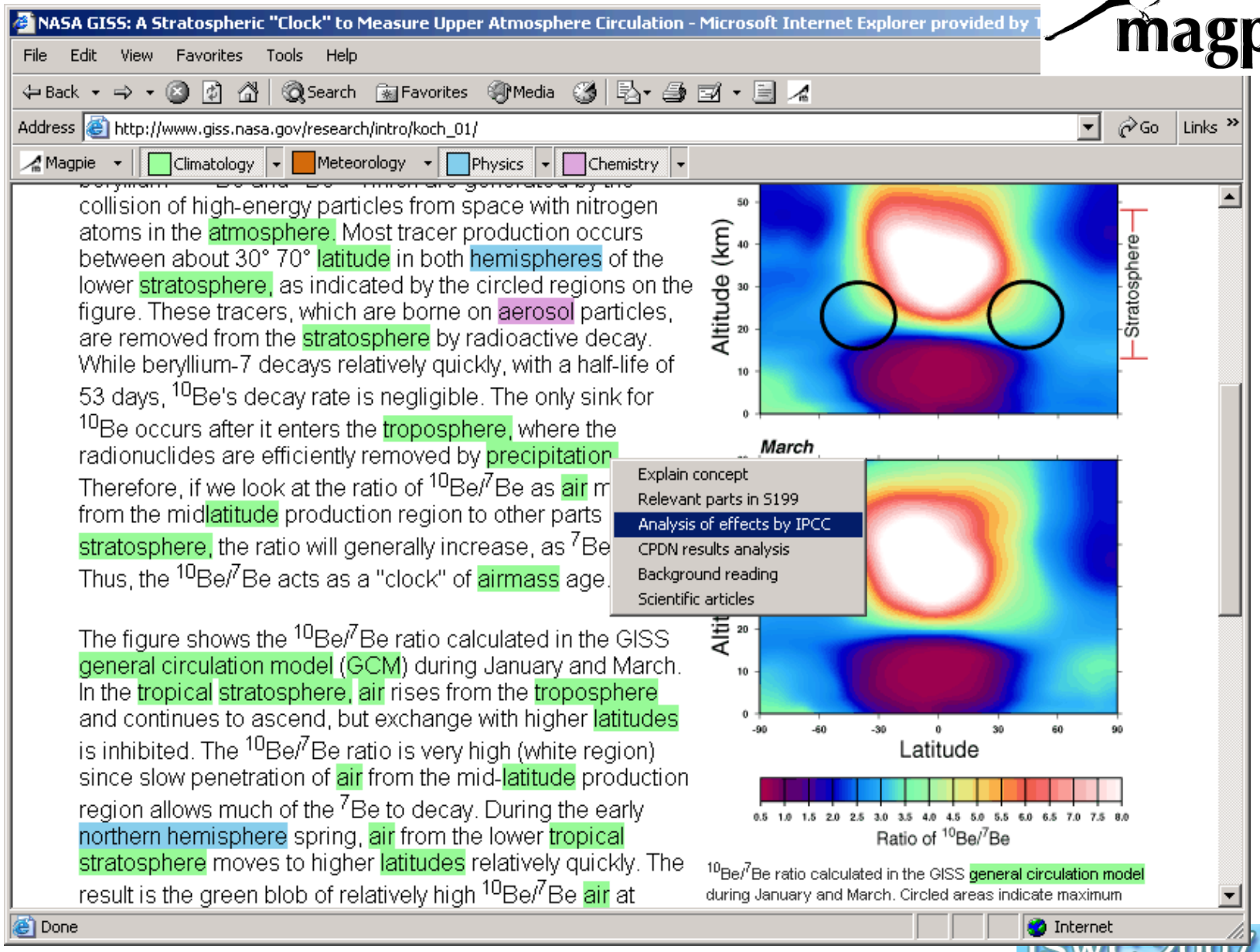
(note: this is not a real legal document!)

ISWC 2007

The University Of Sheffield.

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)



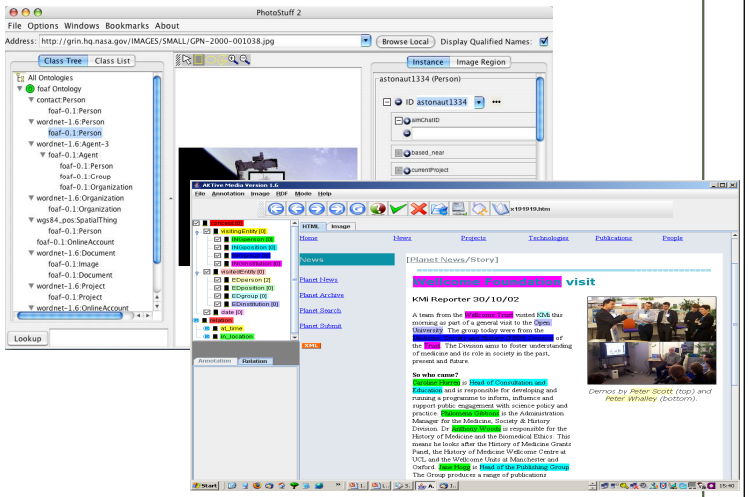
The University Of Sheffield.

Input &Output

- Input to the KA 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)

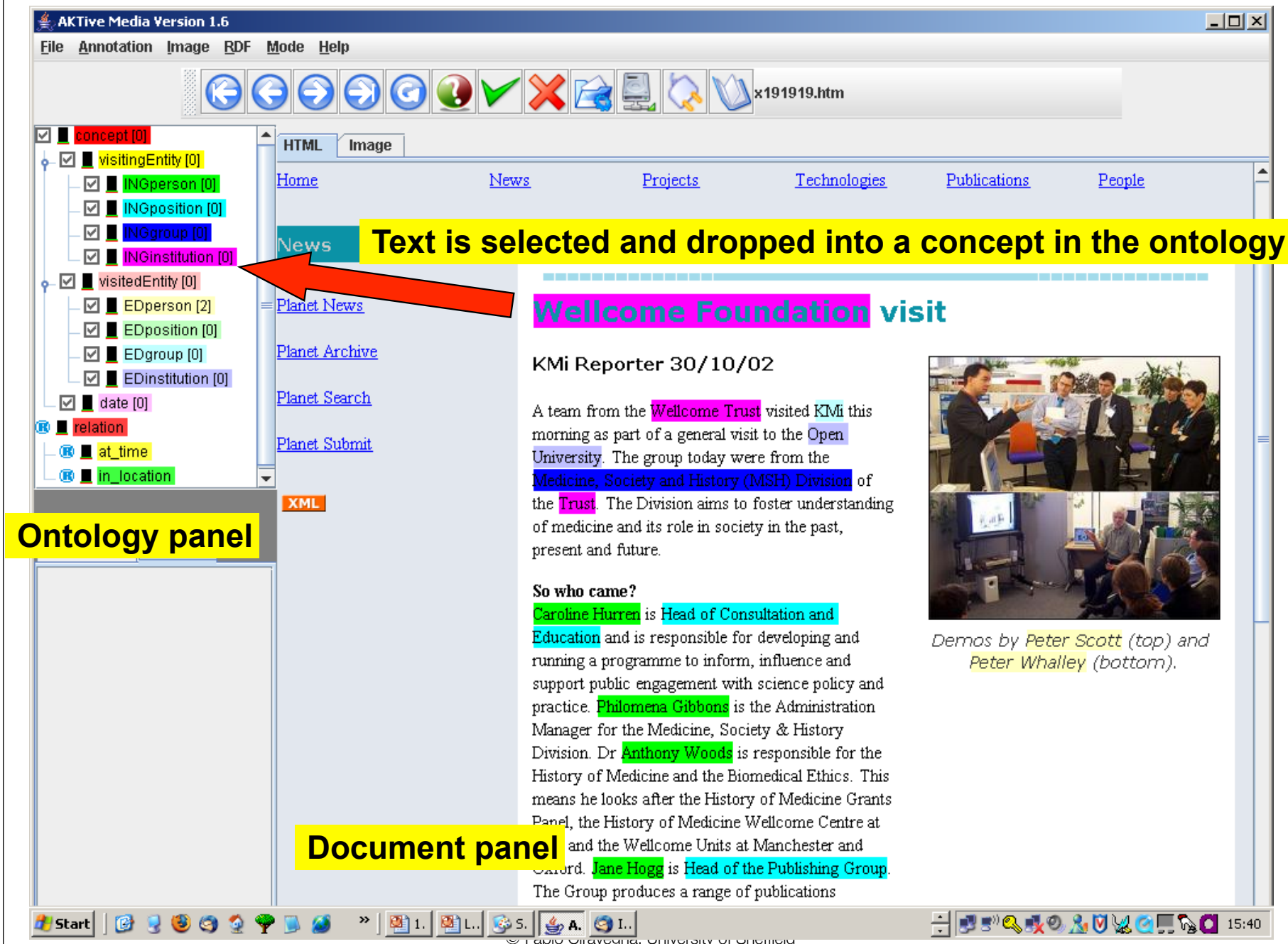


The University Of Sheffield.

AktiveMedia

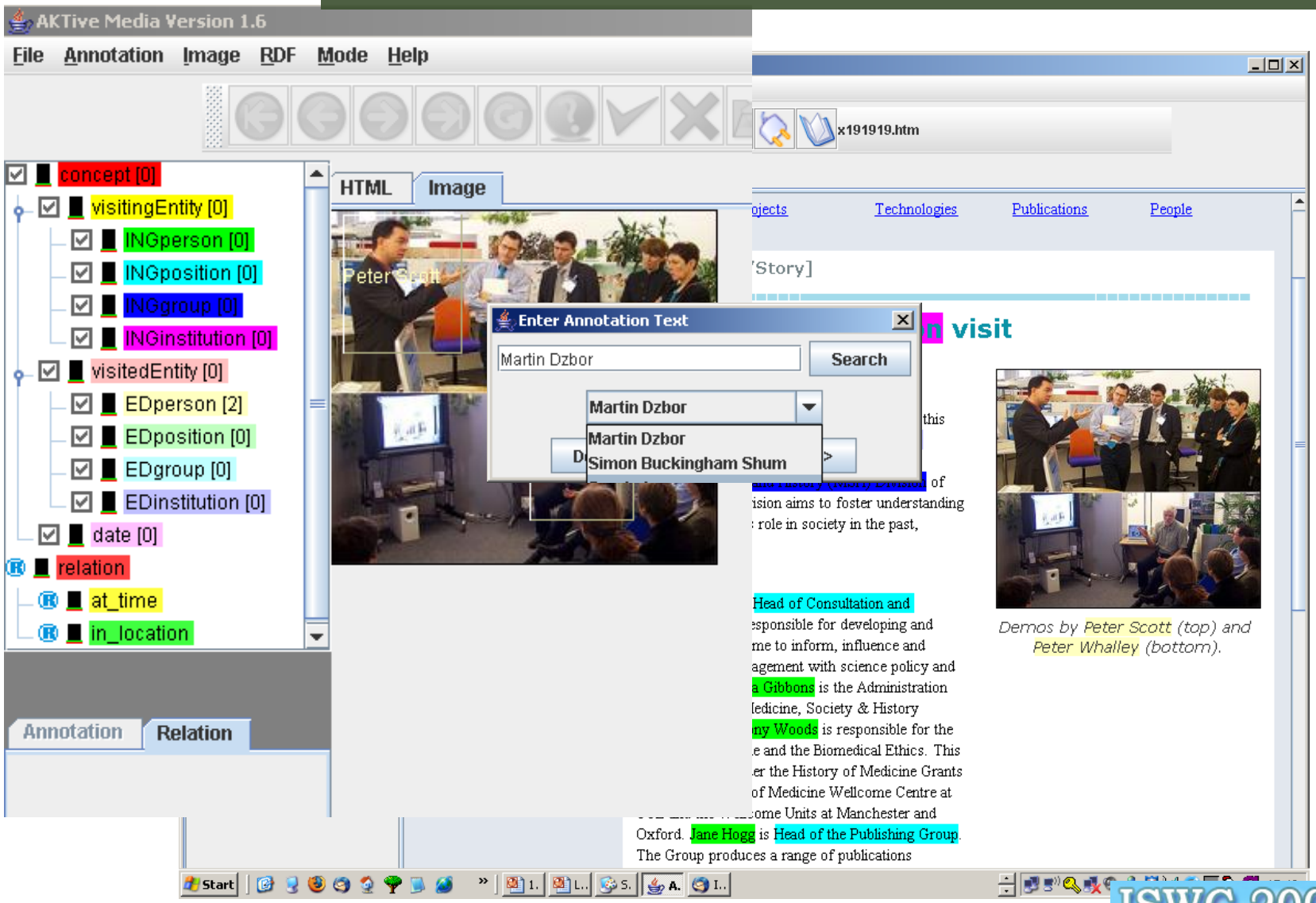
- Enables semi-automatic annotation across texts and images
- The interface enables
  - HTML editing
  - Annotation of documents in RDF based on an OWL ontology
- Types of annotations
  - Concepts / Relations
- SW: Annotation:
  - Selection of concept/relation and highlighting of text is the way in which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

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

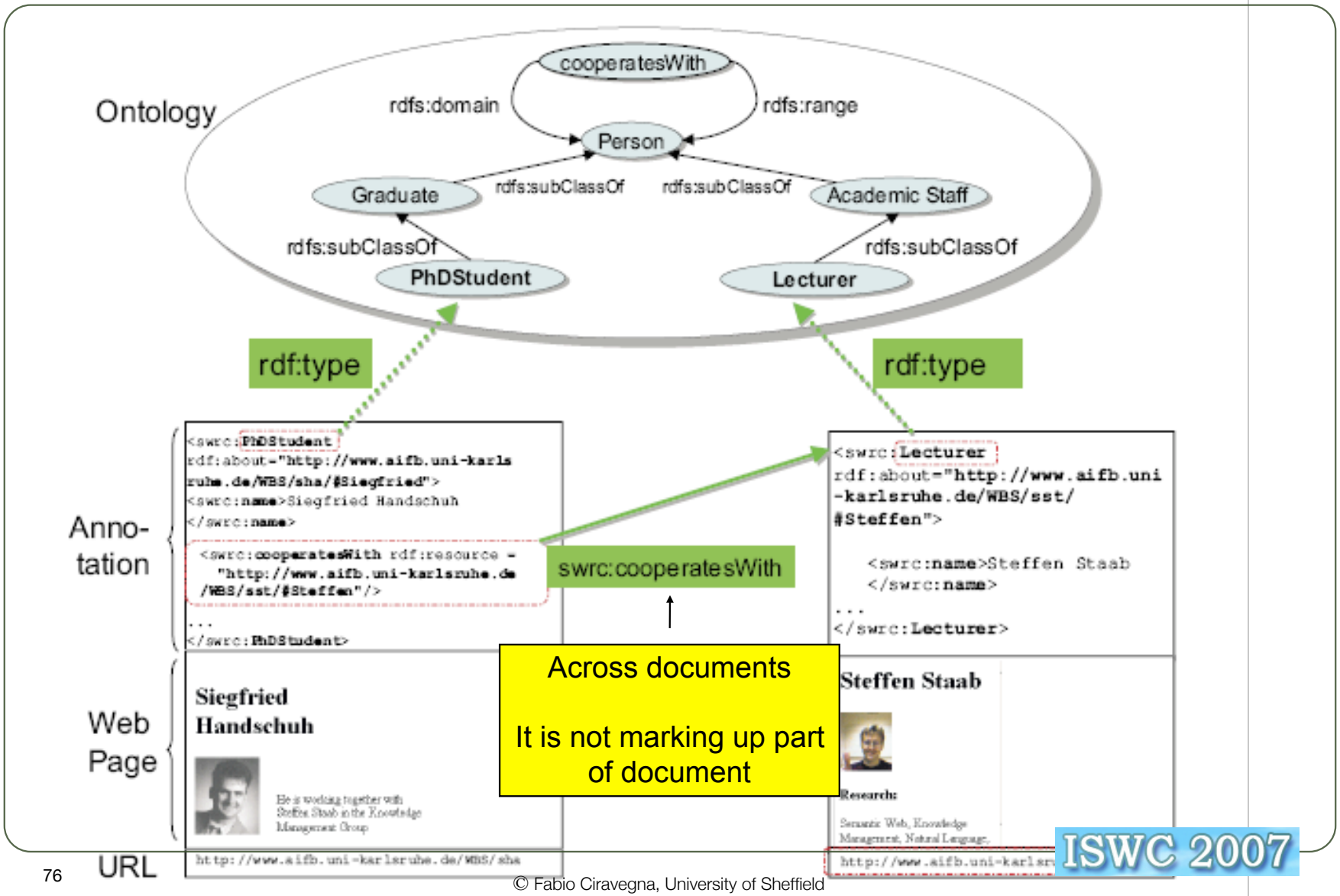


The University Of Sheffield.

Contextual Annotation of Images and Text



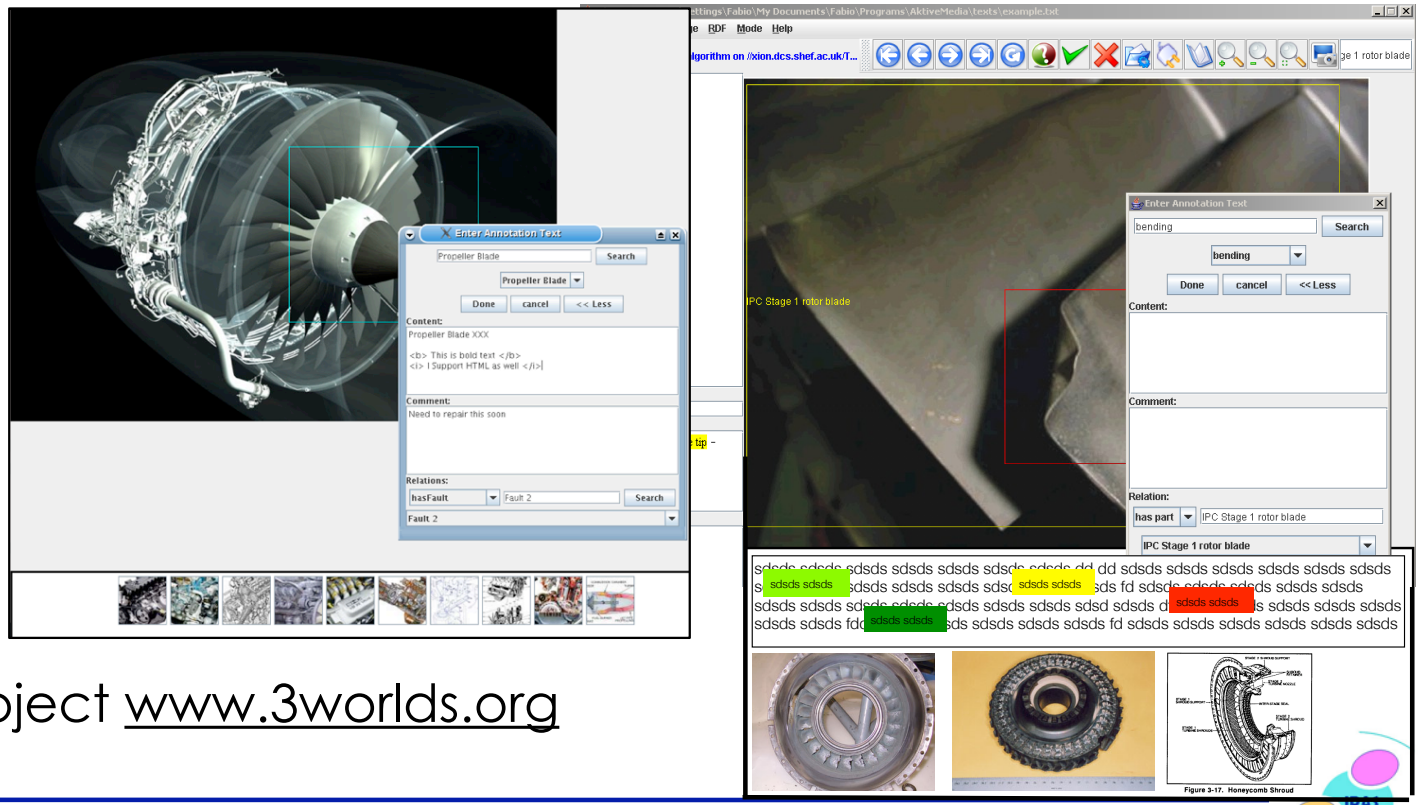
Annotating across documents (CREAM, 2001)



The University Of Sheffield.

Example of Application

- Annotation of compound documents for documenting the overhaul of a jet engine



IPAS project www.3worlds.org

Manual Annotation

A real world application
to knowledge management

ISWC 2007

The knowledge capture tension

- Different departments need specific methods and styles to capture knowledge
  - Justified by internal workflows and tradition
- Rest of company needs sharing in ways suitable to reuse
  - The way information is captured is not necessarily the most appropriate

User-centred Capture in IPAS

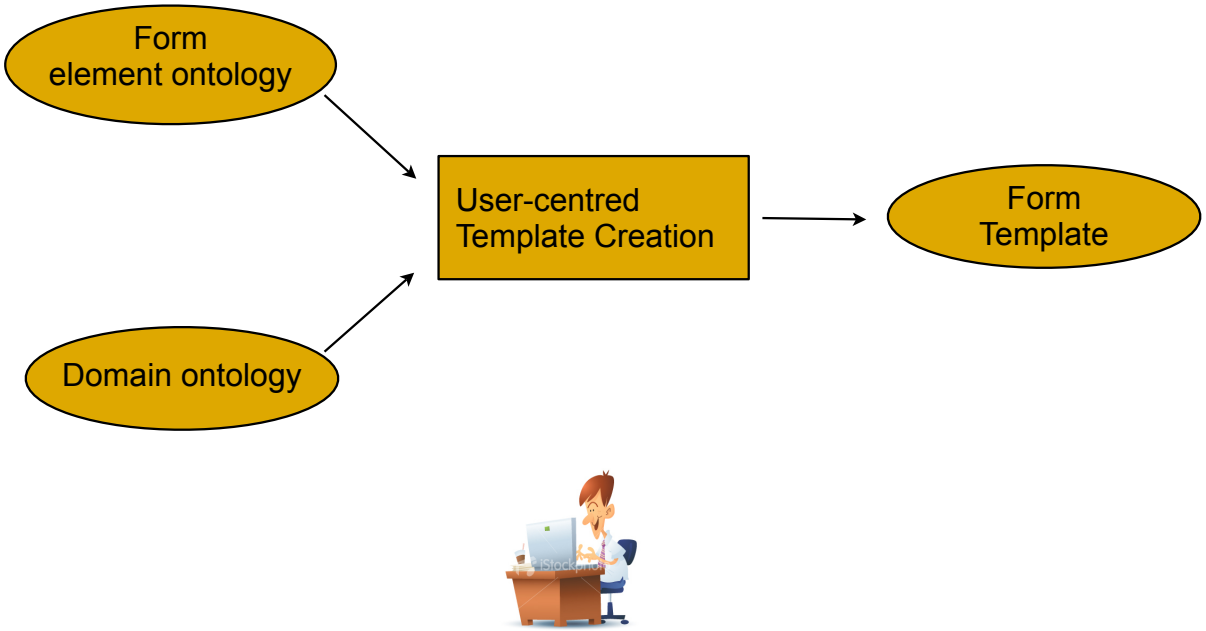
- Currently: single departments establish Word or Excel templates to capture knowledge
  - Knowledge is unstructured
    - It requires effort (e.g. Information Extraction) to extract and share knowledge
- K-Forms
  - Easy user-driven creation of Web based forms to capture knowledge
    - Items in forms are manually connected to company's ontology
      - enables sharing and integration with other knowledge sources
    - knowledge is immediately available for search, sharing and reuse
      - based on k-search
    - capture equivalent to that done by IE, but with user in the loop
    - no additional effort required by user wrt present methodology

Forms

- Enables easy definitions of knowledge capturing applications
- Users define Web based forms visually using a Web browser
- Forms, fields, type of information, possible values, etc.
- Connects form fields to existing ontology or database schema
- Applications can be distributed (via intranet), or local to computer (e.g. laptop)
- Upload of data onto central database in second time (if used in local)
- Real world application to Module Condition Reports at Rolls-Royce plc

Capturing with K-Forms

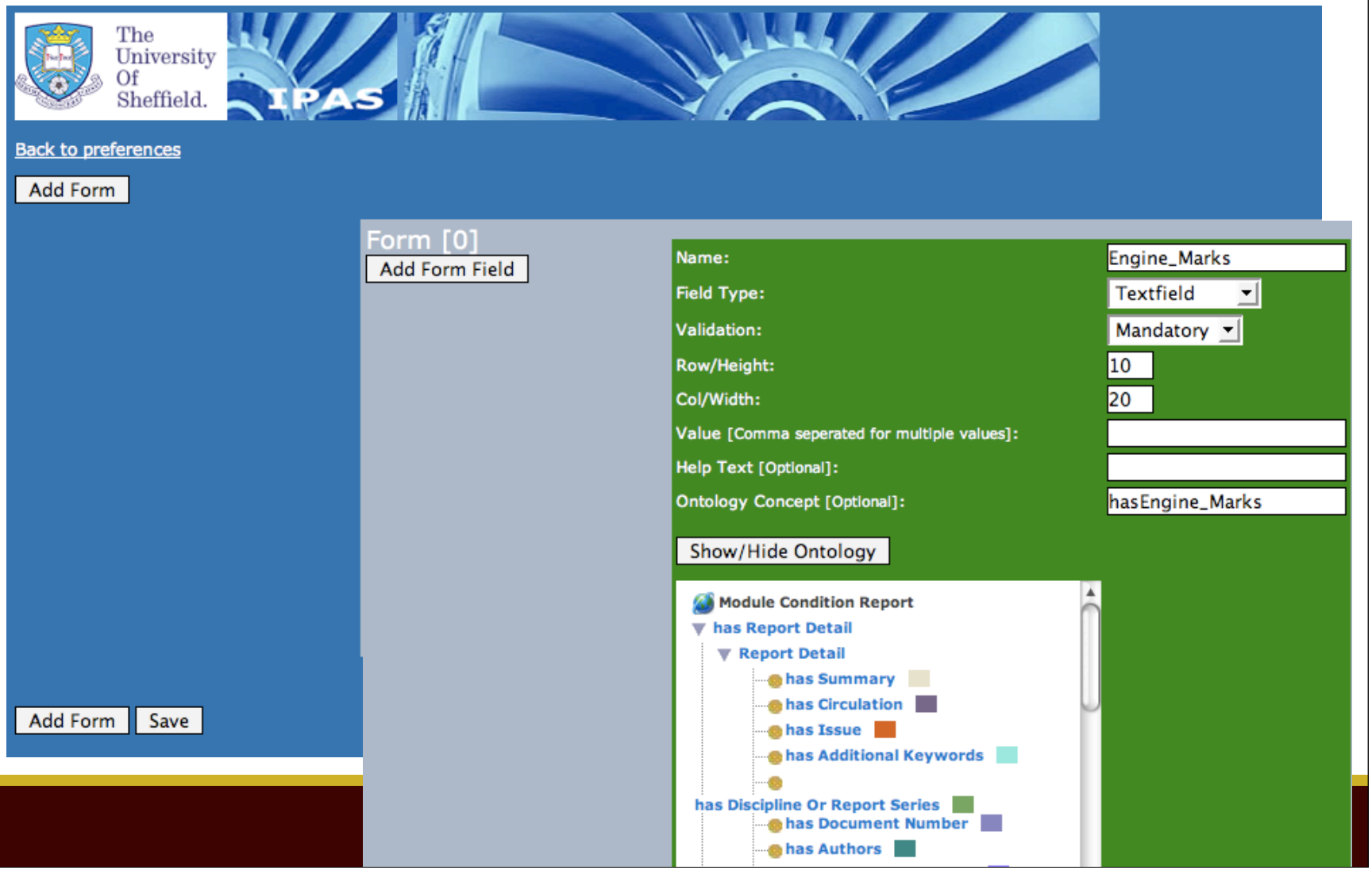
1. Template Design Phase



User: template designer

k-now.co.uk

Easy creation of Web based forms

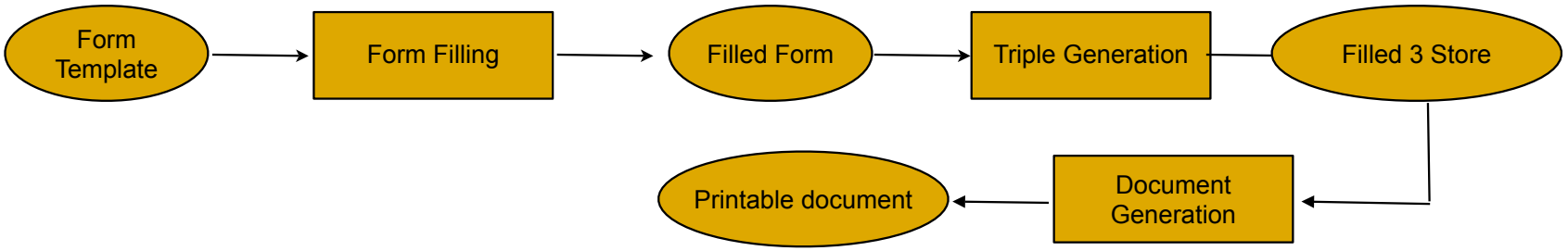


Forms: Features of template creation

- Easy creation and release of form templates over Intranet and in local
- Automatic generation of final document template for each new template
- Based on:
  - Use of ontology of potential form elements
    - Templates composed using ontology directives compiled using a graphical interface
      - All declarative information
    - Ontology is
      - either created automatically around forms and fields
      - or form fields are mapped to ontology concepts and relations

Capturing with K-Forms (2)

2. Knowledge Capture Phase



User: Domain Expert

Forms: Features of Knowledge Capturing

- When form is released users receive forms to fill
  - Easy capture in local (no intranet connection)
  - Easy upload to central repository
- Final document automatically generated
  - Can be read and printed and sent by email
- Knowledge immediately searchable with K-Search after uploading to central repository

The University Of Sheffield.

View Data New Report Publish Thank you for using AKTive IPAS Logo

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
- <rdf:RDF>
  - <rdf:Description rdf:about="http://xion.dcs.shef.ac.uk:8080/aktiveipas/ontologies/generic/module#1.1194256780635">
      <j.0:Repair_0_4>1</j.0:Repair_0_4>
      <j.0:text_MD_Mechanism_0_1_0>Cracked</j.0:text_MD_Mechanism_0_1_0>
      <j.0:Benchmark_2_0/>
      <j.0:TV_And_Accepted_0_6/>
      <j.0:text_Mechanism_1_1_1>Other</j.0:text_Mechanism_1_1_1>
      <j.0:Measurement_data_0_4_2>NONE NONE NONE </j.0:Measurement_data_0_4_2>
      <j.0:Measurement_data_2_4_1>Cracks beyond SP50 limit</j.0:Measurement_data_2_4_1>
      <j.0:Inspection_Locations_0_0>Assembly</j.0:Inspection_Locations_0_0>
      <j.0:IPC_Reference>725221-01-060</j.0:IPC_Reference>
      <j.0:Mechanism_0_1_2>on</j.0:Mechanism_0_1_2>
      <j.0:Invoice_Value_Paid_Repair>NONE</j.0:Invoice_Value_Paid_Repair>
      <j.0:Operator>SIA</j.0:Operator>
      <j.0:SAP_Identification_Number>TR0010708-725221-01-060</j.0:SAP_Identification_Number>
      <j.0:Condition_Details_Pre_Clean_0_3_1/>
      <j.0:text_MD_Mechanism_2_1_1>Other</j.0:text_MD_Mechanism_2_1_1>
      <j.0:text_MD_Mechanism_1_1_0>Corrosion</j.0:text_MD_Mechanism_1_1_0>
      <j.0:Measurement_data_1_4_2/>
      <j.0:Mechanism_0_2>Loose shank nuts</j.0:Mechanism_0_2>
      <j.0:Mechanism_3_2/>
      <j.0:Condition_Pre_Clean_0_2_2>Light frettage to NGV location rails</j.0:Condition_Pre_Clean_0_2_2>
      <j.0:Accept_As_Is_1_5/>
```


moval. It was the engine's

ıal life at the shop visit was

nner race for impact

The University Of Sheffield.

Annotations: Where From?

- SW relies on document annotation
  - Current state of art requires manual annotation
- Manual Annotation for the global Web (Internet)
  - Very few people will annotate web pages by hand
  - What if they did?
    - Isn't the web based on hype?
      - Do people really need to publish their girlfriend photos?

The University Of Sheffield.

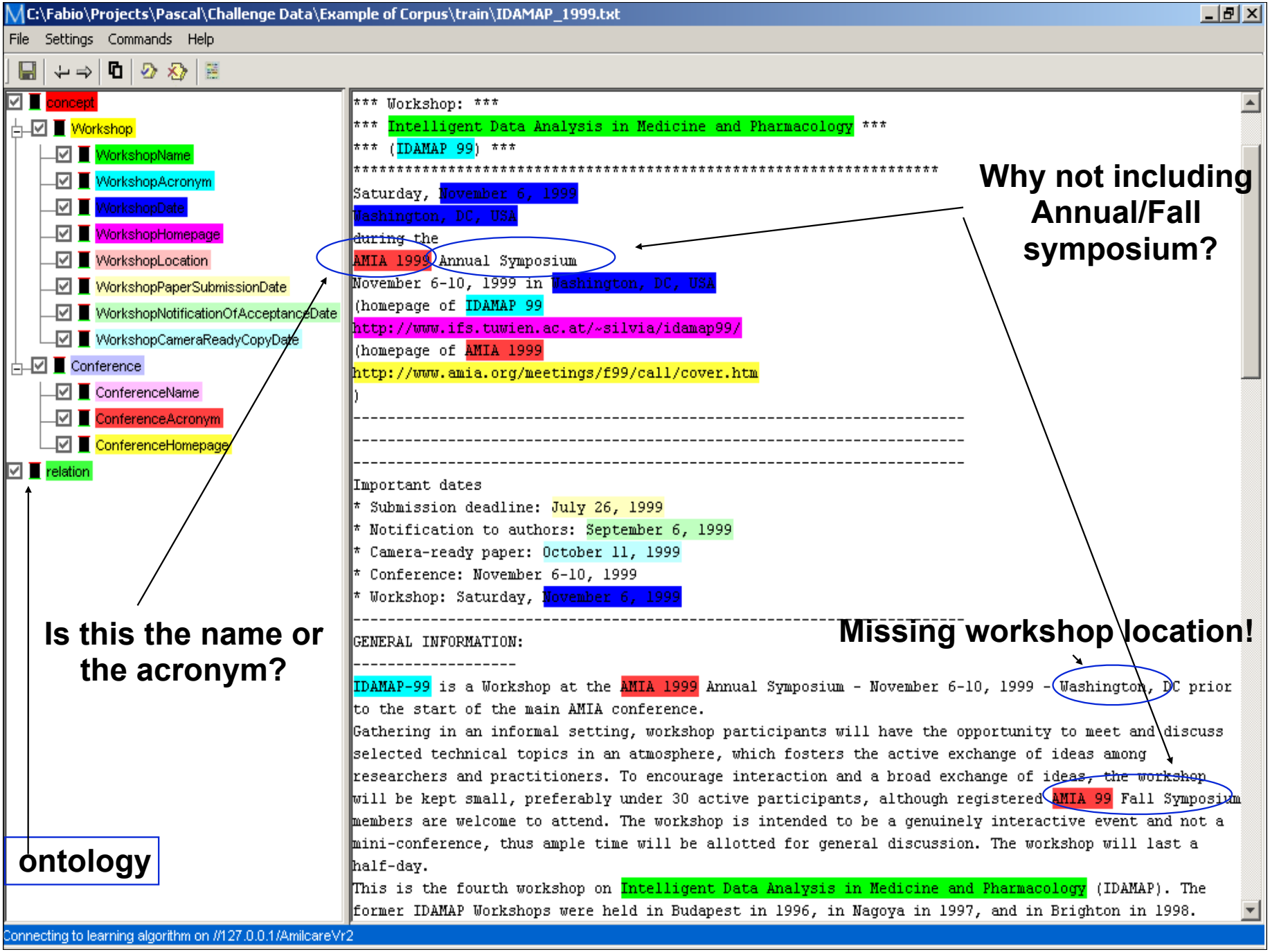
Manual Annotation (1)

- Expensive/time consuming/difficult
  - Chicken-egg problem
  - If it adds time to page editing, users will not do it unless there is really something for them
    - Usefulness and hype
- Inefficient and never ending
  - Every new document needs to be annotated
- Difficult
  - if two people annotate the same documents have 15-30/100 probabilities to annotate them differently
    - Risk is that the same information is annotated differently
      - Disagreement between annotators means data sparsity
      - Information becomes difficult to retrieve

The University Of Sheffield.

An Example

- 10 annotators
- Emails about workshop announcements
  - Name, acronym, date of workshop
  - Name, acronym, URL of associated conference (if any)
  - Submission dates.
- 15% inter-annotator disagreement
  - Especially on name of conference/workshop


C:\Fabio\Projects\Pascal\Challenge Data\Example of Corpus\train\IDAMAP_1999.txt
File Settings Commands Help
concept
Workshop
WorkshopName
WorkshopAcronym
WorkshopDate
WorkshopHomepage
WorkshopLocation
WorkshopPaperSubmissionDate
WorkshopNotificationOfAcceptanceDate
WorkshopCameraReadyCopyDate
Conference
ConferenceName
ConferenceAcronym
ConferenceHomepage
relation
*** Workshop: ***
*** Intelligent Data Analysis in Medicine and Pharmacology ***
*** (IDAMAP 99) ***
Saturday, November 6, 1999
Washington, DC, USA
during the
AMIA 1999 Annual Symposium
November 6-10, 1999 in Washington, DC, USA
(homepage of IDAMAP 99
http://www.ifs.tuwien.ac.at/~silvia/idamap99/
(homepage of AMIA 1999
http://www.amia.org/meetings/f99/call/cover.htm
)
Important dates
* Submission deadline: July 26, 1999
* Notification to authors: September 6, 1999
* Camera-ready paper: October 11, 1999
* Conference: November 6-10, 1999
* Workshop: Saturday, November 6, 1999
GENERAL INFORMATION:
IDAMAP-99 is a Workshop at the AMIA 1999 Annual Symposium - November 6-10, 1999 - Washington, DC prior to the start of the main AMIA conference.
Gathering in an informal setting, workshop participants will have the opportunity to meet and discuss selected technical topics in an atmosphere, which fosters the active exchange of ideas among researchers and practitioners. To encourage interaction and a broad exchange of ideas, the workshop will be kept small, preferably under 30 active participants, although registered AMIA 99 Fall Symposium members are welcome to attend. The workshop is intended to be a genuinely interactive event and not a mini-conference, thus ample time will be allotted for general discussion. The workshop will last a half-day.
This is the fourth workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP). The former IDAMAP Workshops were held in Budapest in 1996, in Nagoya in 1997, and in Brighton in 1998.
Connecting to learning algorithm on //127.0.0.1/AmilcareVr2
Why not including Annual/Fall symposium?
Is this the name or the acronym?
Missing workshop location!
ontology

The University Of Sheffield.

Problems in the example

- The previous example contains
  - Three doubtful cases (conference name/acronym)
  - One mistake
- It was annotated by two people and a third one checked their annotations
  - It was in a corpus of hundreds of documents!

ISWC 2007

The University Of Sheffield.

Problems with Manual Annotation (2)

- Tedious & Tiring
  - Error prone
- Legacy with the past
  - Ontologies are living objects, new version produced
    - Which version of the ontology is used for annotation?
- Dispersed information
  - Annotation largely unfeasible for large diverse repositories
    - E.g. a Web site
      - Department of CS of the University of Southampton: 1,600 pages
    - How many relevant ontologies are there for that department?

ISWC 2007

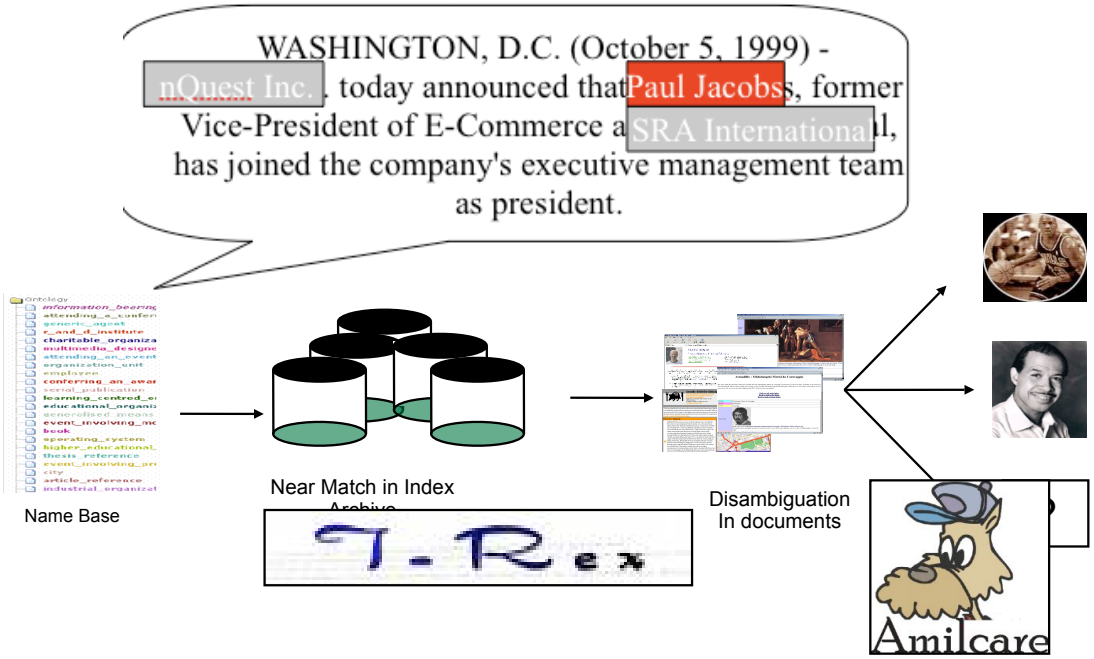
Problems with Manual Annotation (3)

- How many annotation schemas?
  - The Semantic Web is expected to be composed of
    - [Many] small ontological components [Hendler 2001] will be created, mainly related to different domain and applications
    - University of Sheffield web site:
      - What ontology for annotation?
        - Universities/Education, Research life, Scientific Papers,
        - Sport, computer network organization….
        - You name what…

The University Of Sheffield.

Annotation for use…

- If annotation is to be chosen by author/owner
  - Selection of Annotation Schema may reflect world model of the creator, not of the user
    - E.g. education is the main goal of the university, so the central Uni will probably choose an ontology on Education
    - Most of my time is actually devoted to research
    - Most of my colleagues look for scientific information on our web site
    - To us, Uni's annotation would be largely unuseful
    - Question:
      - Who (and how!) is going to introduce the annotation for us?
      - Where is the annotation to be inserted?



Automating Annotation

The University Of Sheffield.

Annotation Engines

- Manual document annotation is still largely expected to be the main SW vehicle creation
  - Especially for trusted environment (e.g. within a company) this is expected to provide high quality material
- But manual annotation is largely unfeasible over large scale
  - Unavailability of users
  - Risks related to incompetence or spamming
  - Large amount of legacy data
- Automatic annotation
  - To help manual annotation OR
  - To replace human annotators (e.g. on legacy data)

ISWC 2007

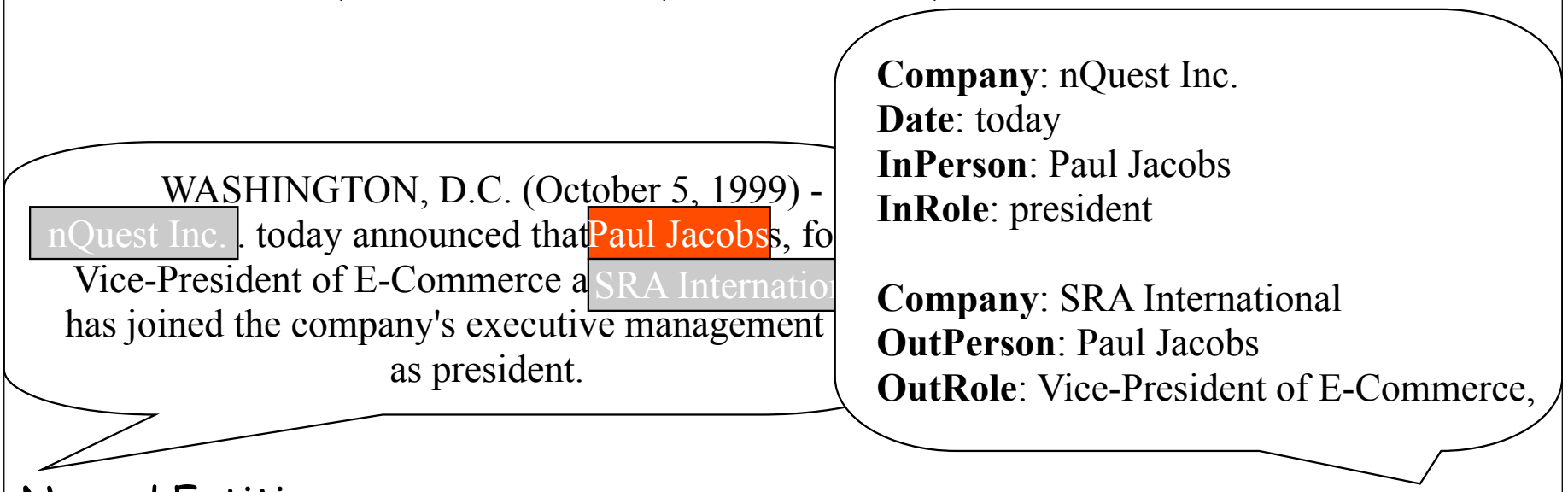
The University Of Sheffield.

Tasks for KA: Extraction

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

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.



The University Of Sheffield.

Named Entity Recognition

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

The University Of Sheffield.

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

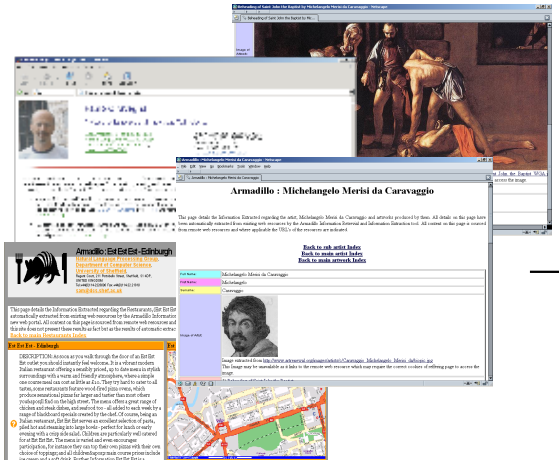
The University Of Sheffield.

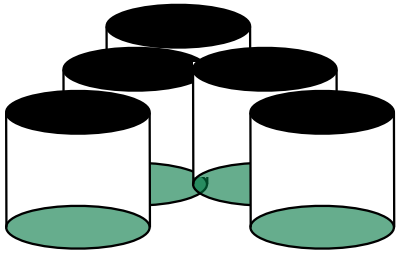
Large Scale NER&C

- 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

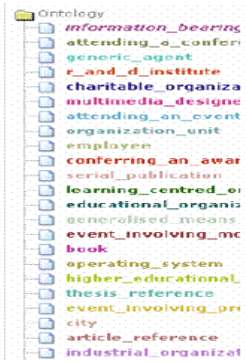
- Document Indexing as in Search Engines



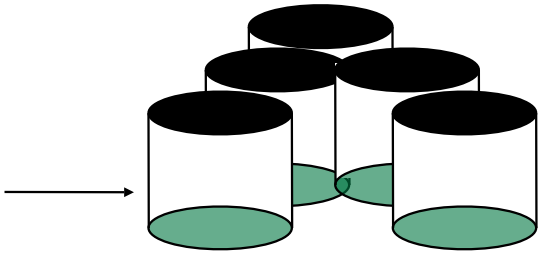


Distributed Index Archive
(keywords)

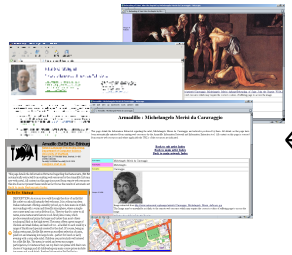
Known Name Recognition



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

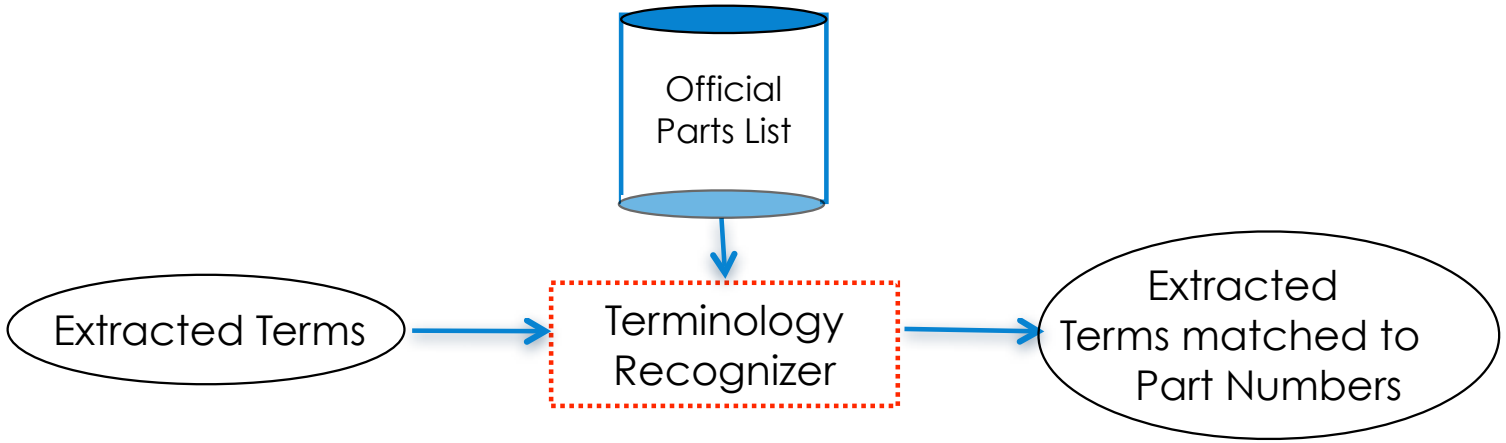
The University Of Sheffield.

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

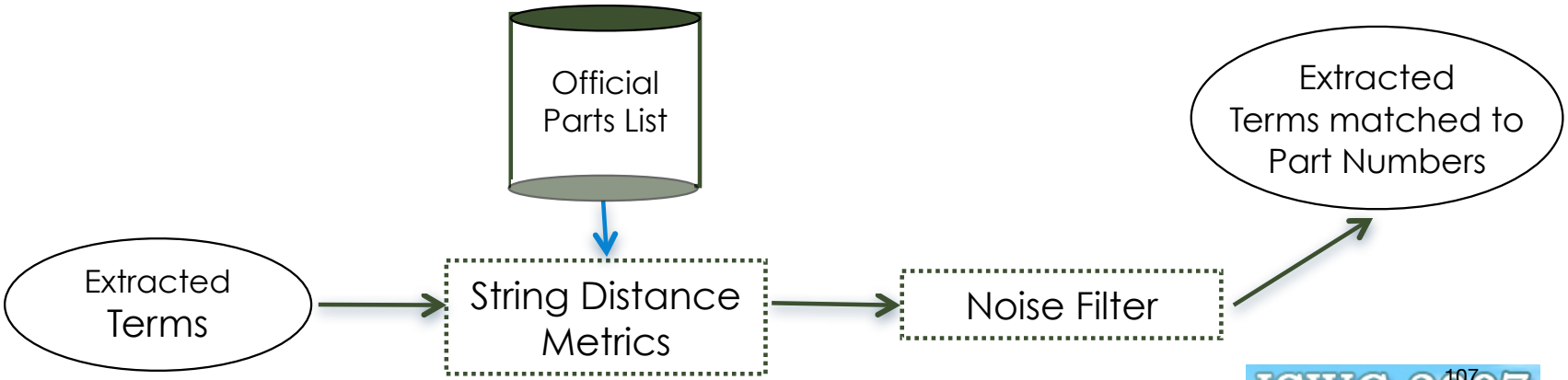
Terminology Recognition

- 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)



The University Of Sheffield.

More complex IE: event modelling

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

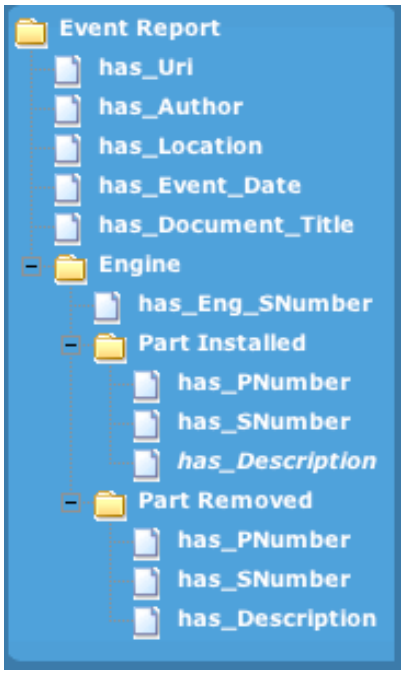
The University Of Sheffield.

Table Field Extraction

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

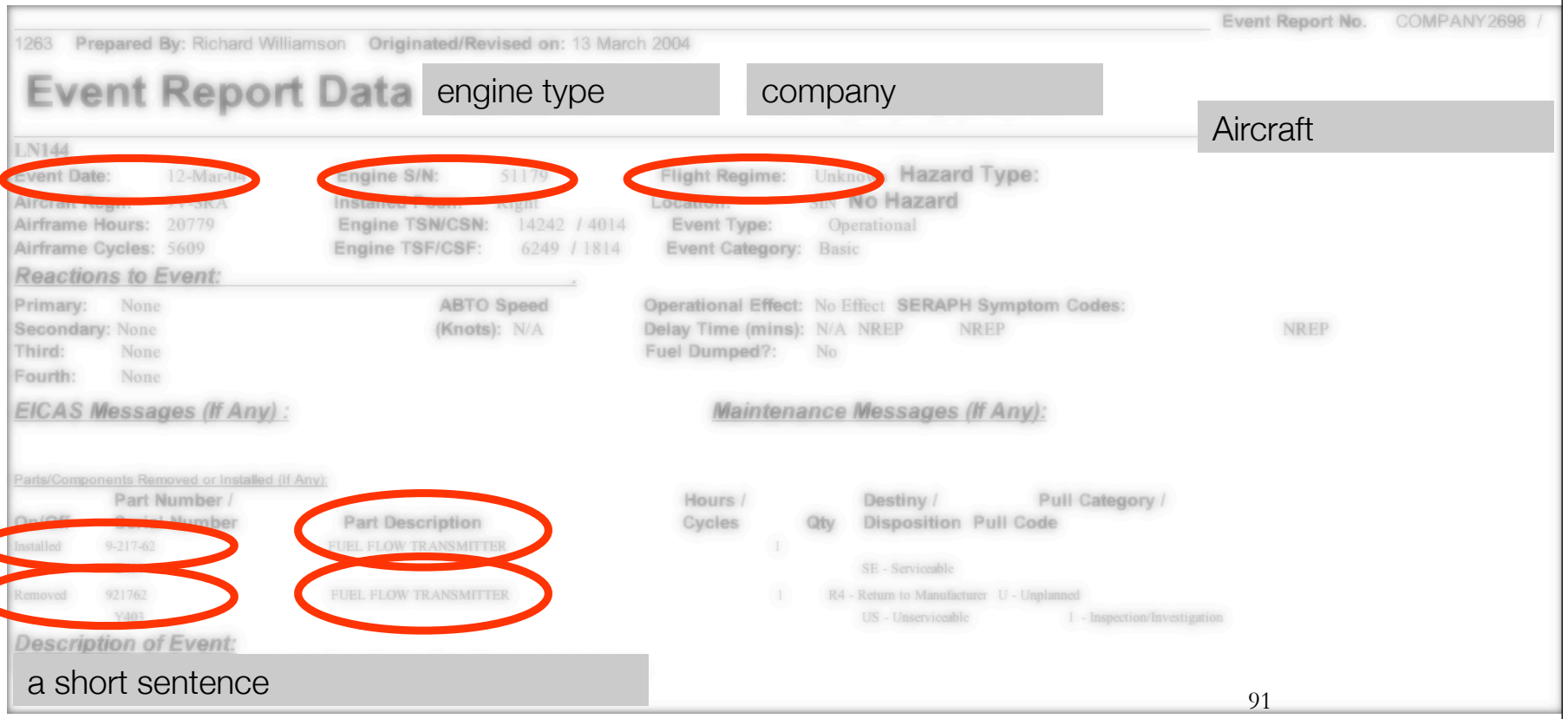
ISWC 2007

An Example of Automatic IE



- Automatic extraction of information from event report
  - 18,000 documents analysed
- Metadata generated according to a simple ontology
- Automatic extraction of metadata and indexing of documents

An Experiment on Event Report for Jet Engines



91

The University Of Sheffield.

Types of tables in Event Reports

module/accessory details			
item	**part number**	**s/n removed**	**s/n installed**
	p39-401revf	04-0721257 tsn/csn: 268/106	04-1012229 tsn/csn:0/0

Part numbers
04-0721257 tsn/csn: 268/106 off
04-1012229 tsn/csn:0/0 on

s/n removed	04-0721257 tsn/csn: 268/106
s/n installed	04-1012229 tsn/csn:0/0

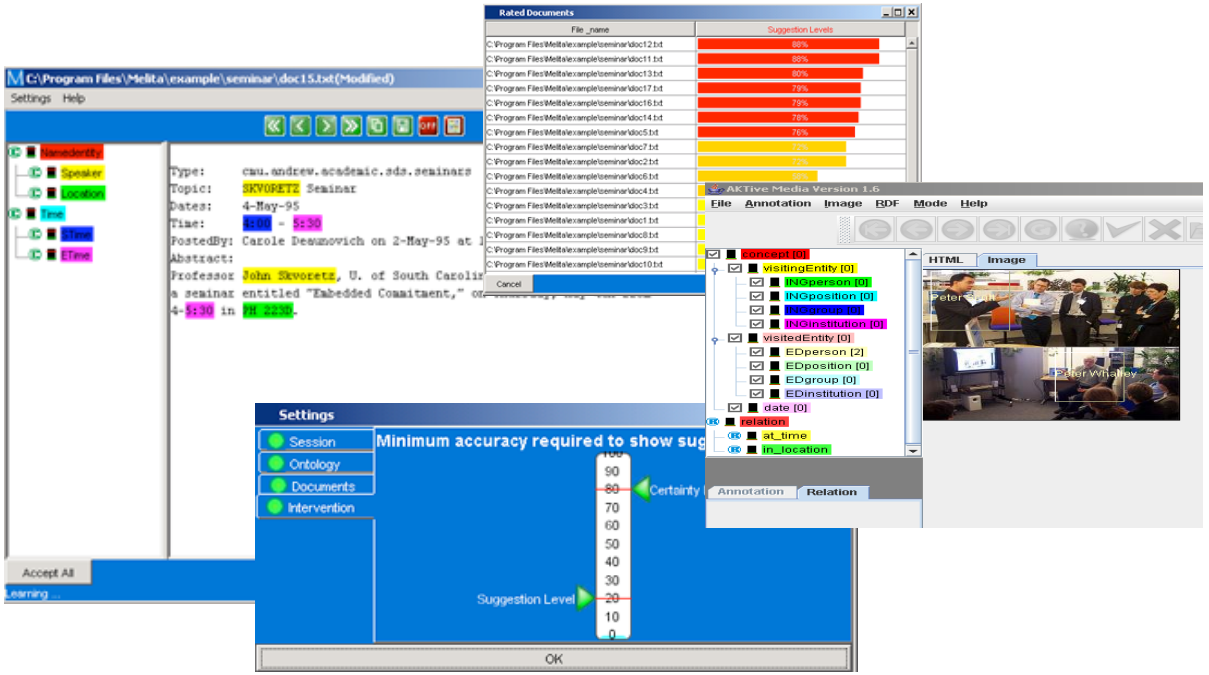
Parts/Components Removed or Installed (If Any):

On/Off	Part Number / Serial Number	Part Description	Hours / Cycles	Qty	Destiny Disposit
Installed	FK30840	TO SB72-C629)	11129 TSN 1954	1	
	RGG12340				
Installed	FK21221 EC092		11652 TSN 2119	1	
Installed	FK30840		11129 TSN 1954	1	
	RGG12501				
Installed	FK30840		11129 TSN 1954	1	
	RGG12208				
Installed	FK30840		11129 TSN 1954	1	
	RGG12391				

Applying information extraction

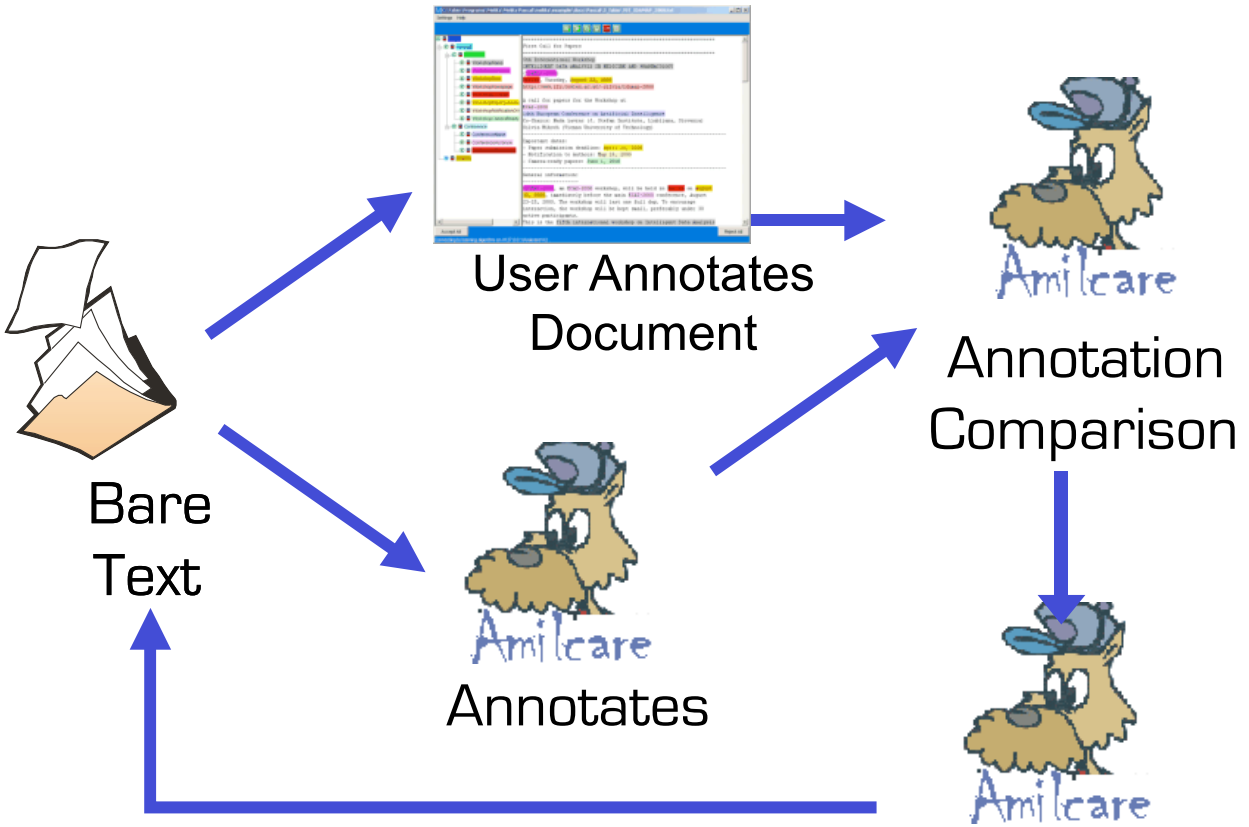
- AktiveMedia to annotate texts
- TRex system (Jiria et al. 2006) to train and extract
  - http://tyne.shef.ac.uk/t-rex/
- IE captures most of the information in tables
  - 99% of the information captured (recall=99)
  - 98% of proposed information is correct (precision=98)

	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**

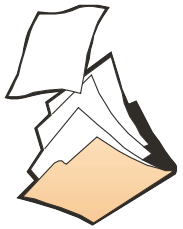


Using IE to Support Manual Annotation

Using IE to support annotation: step 1



Using IE to support annotation: step 2



Bare Text

Annotates

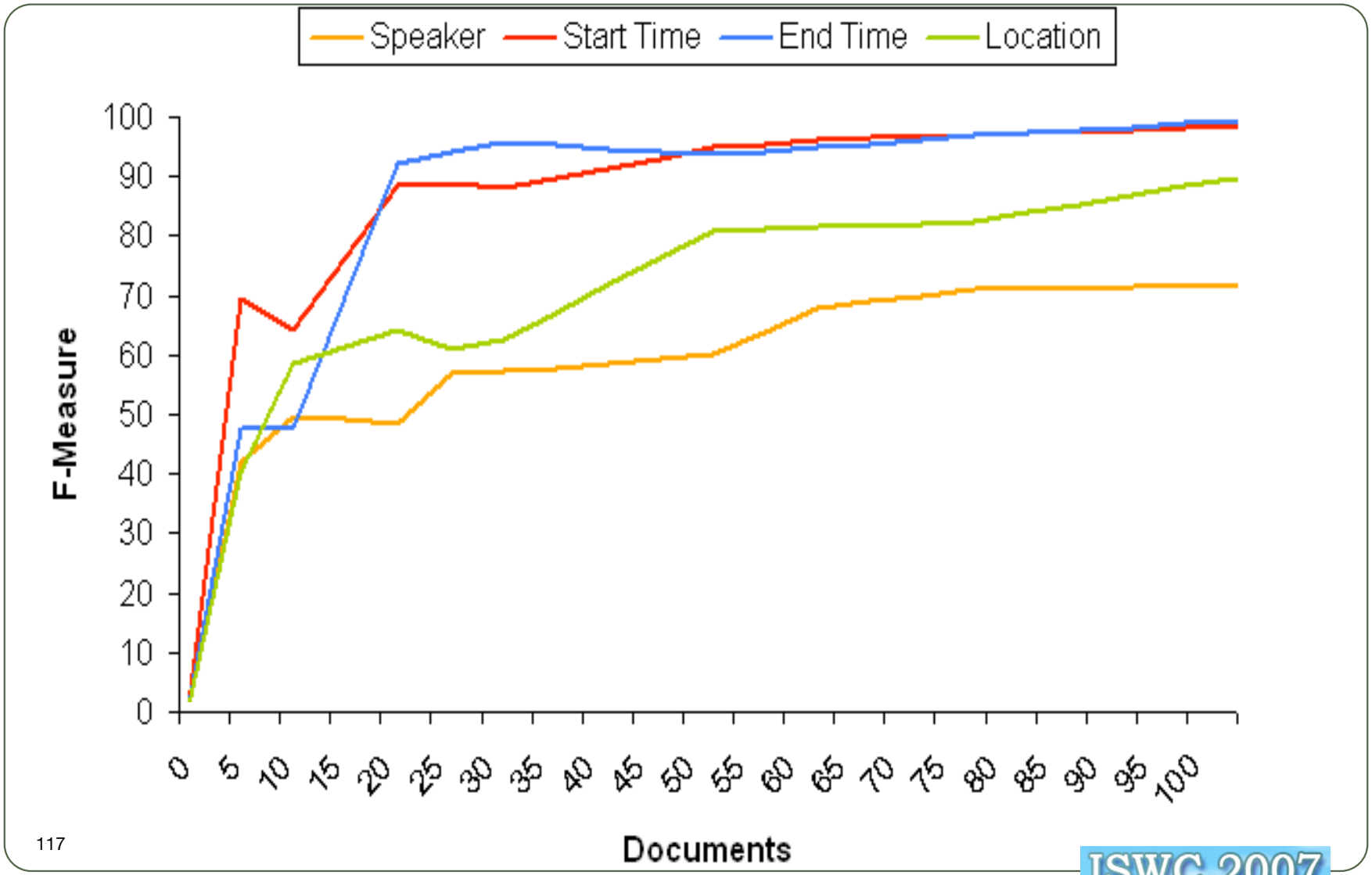
User Corrects

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

The University Of Sheffield.

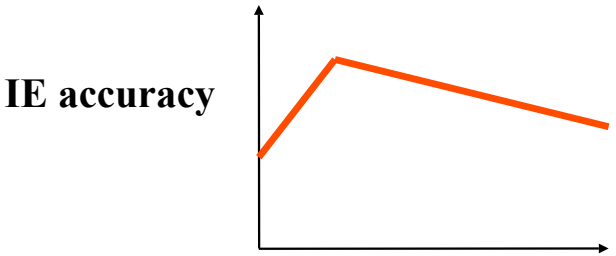
Learning curve

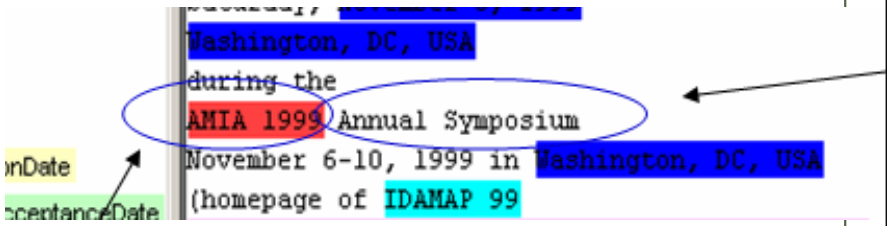


ISWC 2007

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





ISWC 2007

Information Integration

The University Of Sheffield.

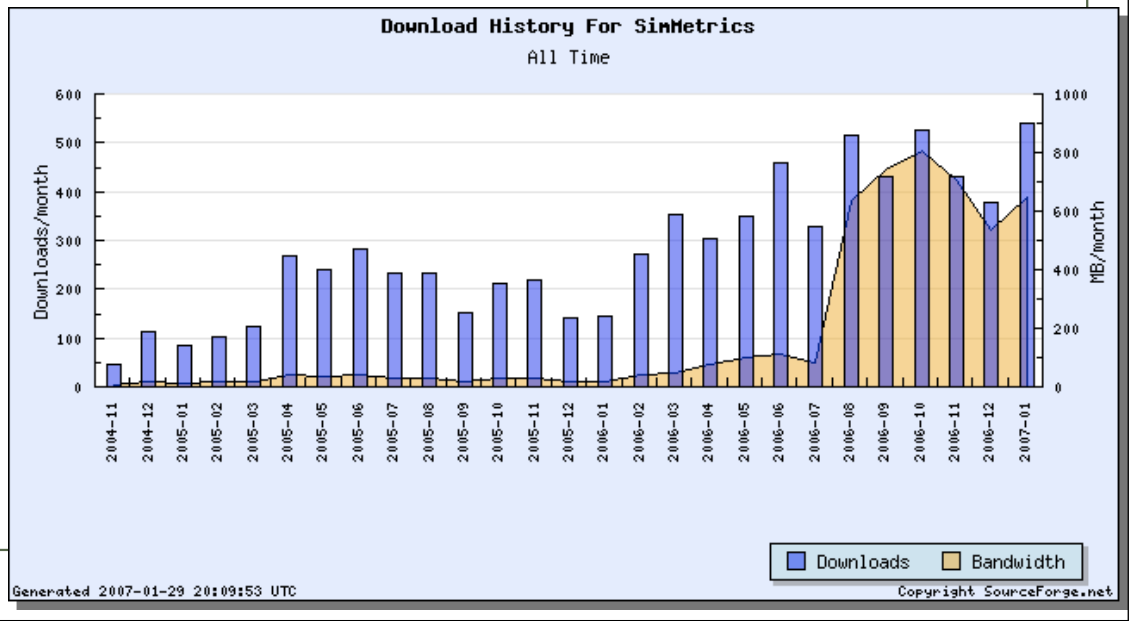
Information Integration

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

The University Of Sheffield.

SimMetrics

- Library of distance metrics released as open source
  - http://sourceforge.net/projects/simmetrics/
  - 7,489 downloads in 2 years
  - Most downloaded distance metrics library on the Web
  - Hundreds of applications



The University Of Sheffield.

Armadillo: Historical Data Mining

Arts & Humanities Research Council

The Marine Society Registers

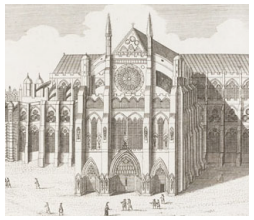
The Westminster Historical Database

Eighteenth Century Fire Insurance Policies

Prerogative Court of Canterbury Wills

The Proceedings of the Old Bailey

AHDS Deposits



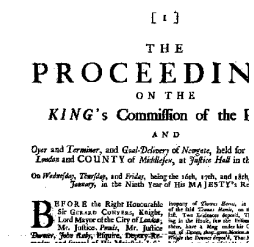
St. Martin's Settlement Exams Index
WESTCAT

Collage image databse
Guildhall Library

Harben's Dictionary of London

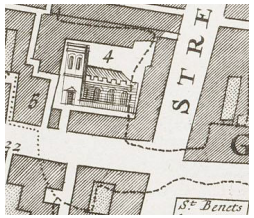
John Strype's "Survey…"

http://www.motco.com



Metropolitan London in the 1690s
IHR

Selected Criminal Records
PRO



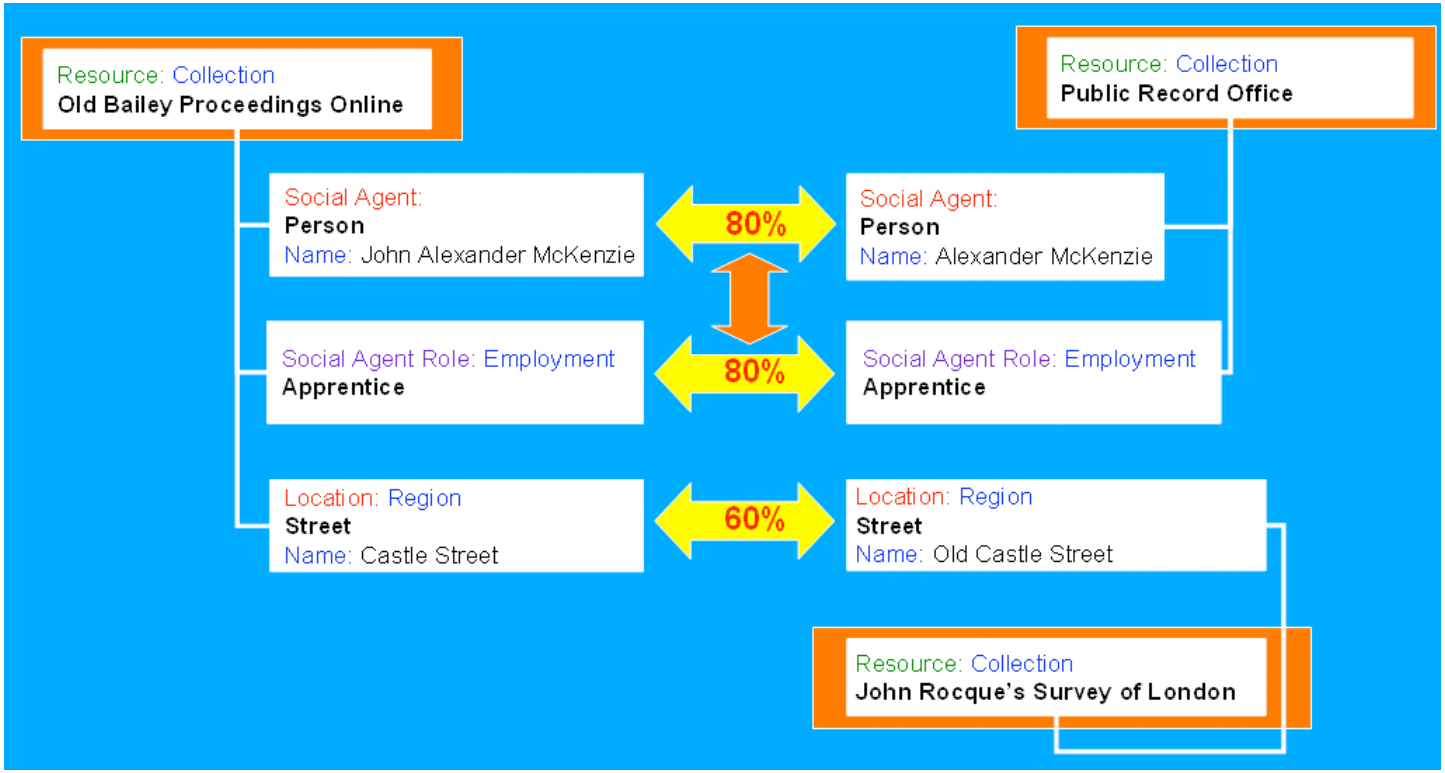
House of Lords Journals
BOPCRIS

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

ISWC 2007

Information Integration

Armadillo: Historical Data Mining



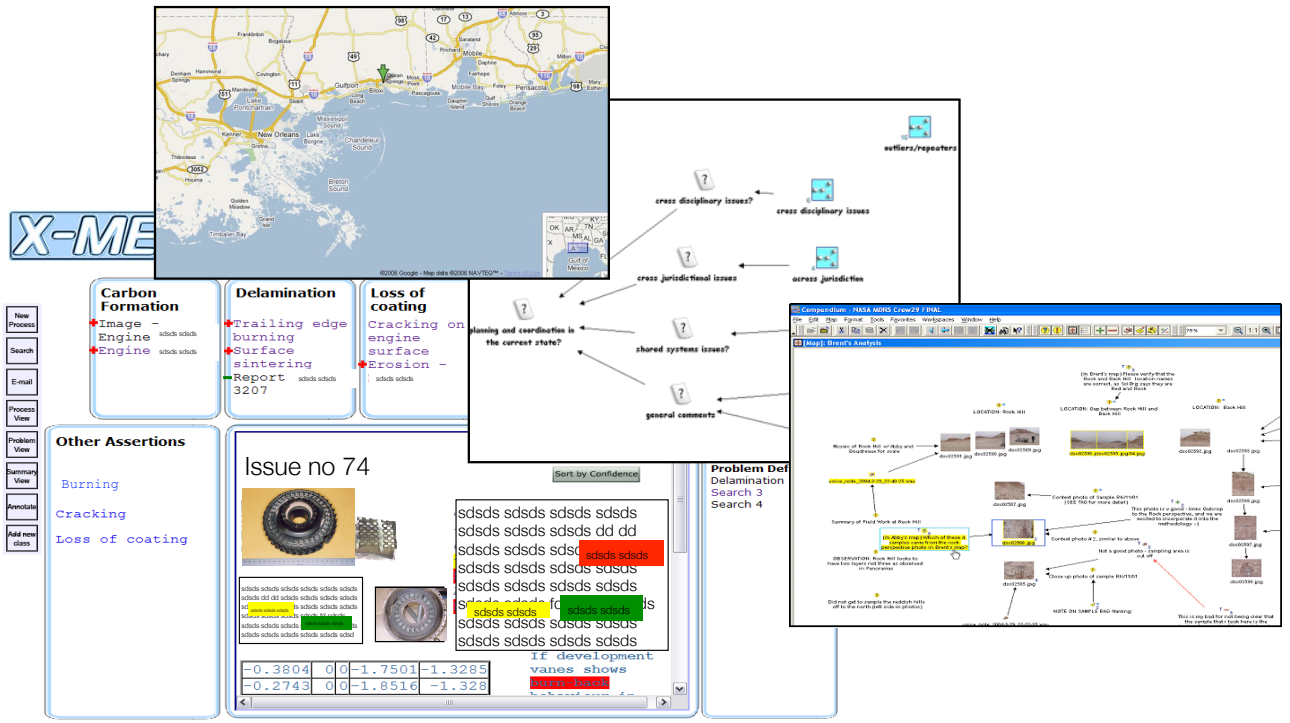
Arts & Humanities Research Council

ISWC 2007

The University Of Sheffield.

Conclusions on KA, Requirements and Tools

- Large scale?
  - Ontologies:
    - large ontologies (up to 10k) with simple tasks (SemTag and Seeker, Kim)
    - small/medium scale (up to 100) with more complex tasks
  - KB: large scale
- Portability: most technology difficult to port without experts (Armadillo, KIM)
  - User input well exploited in human-centred acquisition (e.g. Melita, AktiveMedia)
- Cross-Media: exploited in user centred annotation (e.g. AktiveMedia)
- Background Knowledge
  - Used in AktiveMedia, KIM, SemTag and Armadillo to some extent
  - Uncertainty: some use in Armadillo



Knowledge Sharing and Reuse

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

ISWC 2007

The University Of Sheffield.

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

Requirements for KS&R: Integration

- Large distributed archives require ability to
  - Map distribution of information,
  - Weight every single source
  - Distribute searches carefully;
- Currently: search is performed just in some of the archives, disregarding others that can bring very useful information
- Importance of context and provenance in searching
  - *Sylvester attacked Tweety and Tweety flew away* (Tweety is a tweety hence tweeties fly)
  - *Tweety came back from hospital with a broken wing. Sylvester attacked her. And Tweety could not fly away* (tweety is a tweety hence tweeties do NOT fly ???)

The University Of Sheffield.

Requirements for KS&R: Focussing

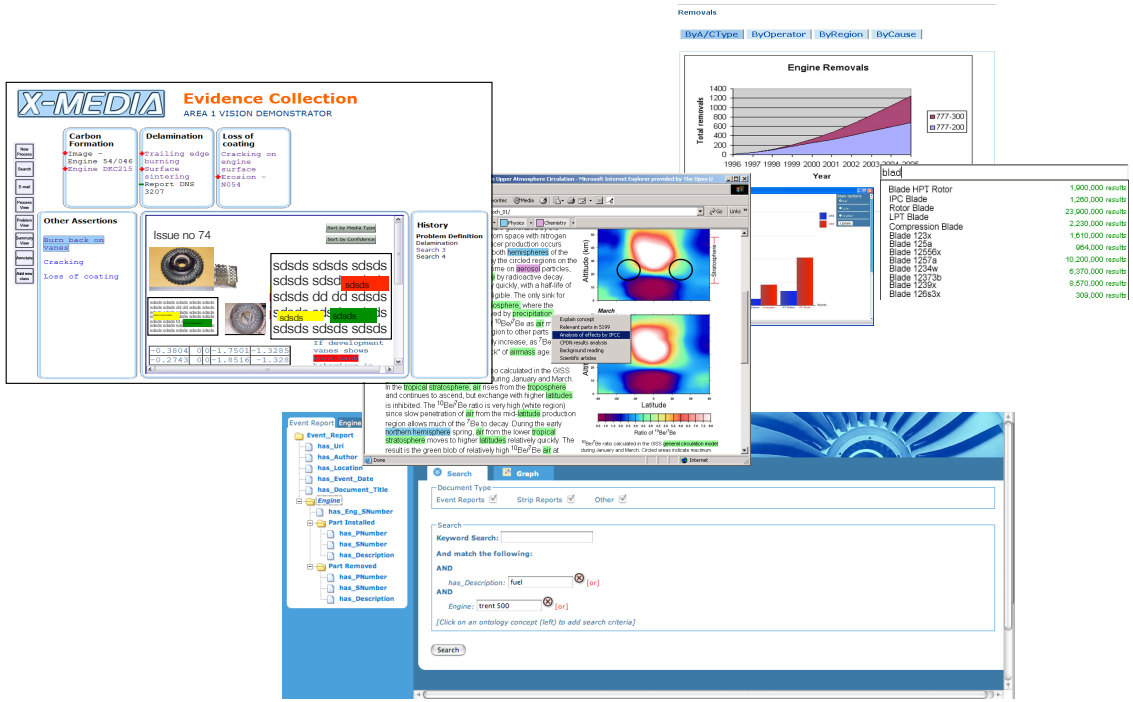
- Managing knowledge becomes more complex and needs powerful focusing methodologies.
- Focus of searching
  - Changes in time and from user to user,
  - Requires a balanced mixture of exploration and searching;
- Focus on what I know: My knowledge as a basis of what to look for
    - My context rather than everyone's context
      - Tell me what I do not know
        - What is that other people know that I do not
        - Tell me more about what I know

ISWC 2007

The University Of Sheffield.

Requirements of KS&R: Focussing (2)

- What is the user interested in:
  - Most frequent phenomena
    - Redundancy-based approaches to KA can work
  - In less frequent phenomena
    - Redundancy-based approaches do not work
  - Domain specific metrics
    - e.g. disruption caused to customers
    - A mix of the two above



SW for Knowledge Sharing and Reuse

The University Of Sheffield.

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

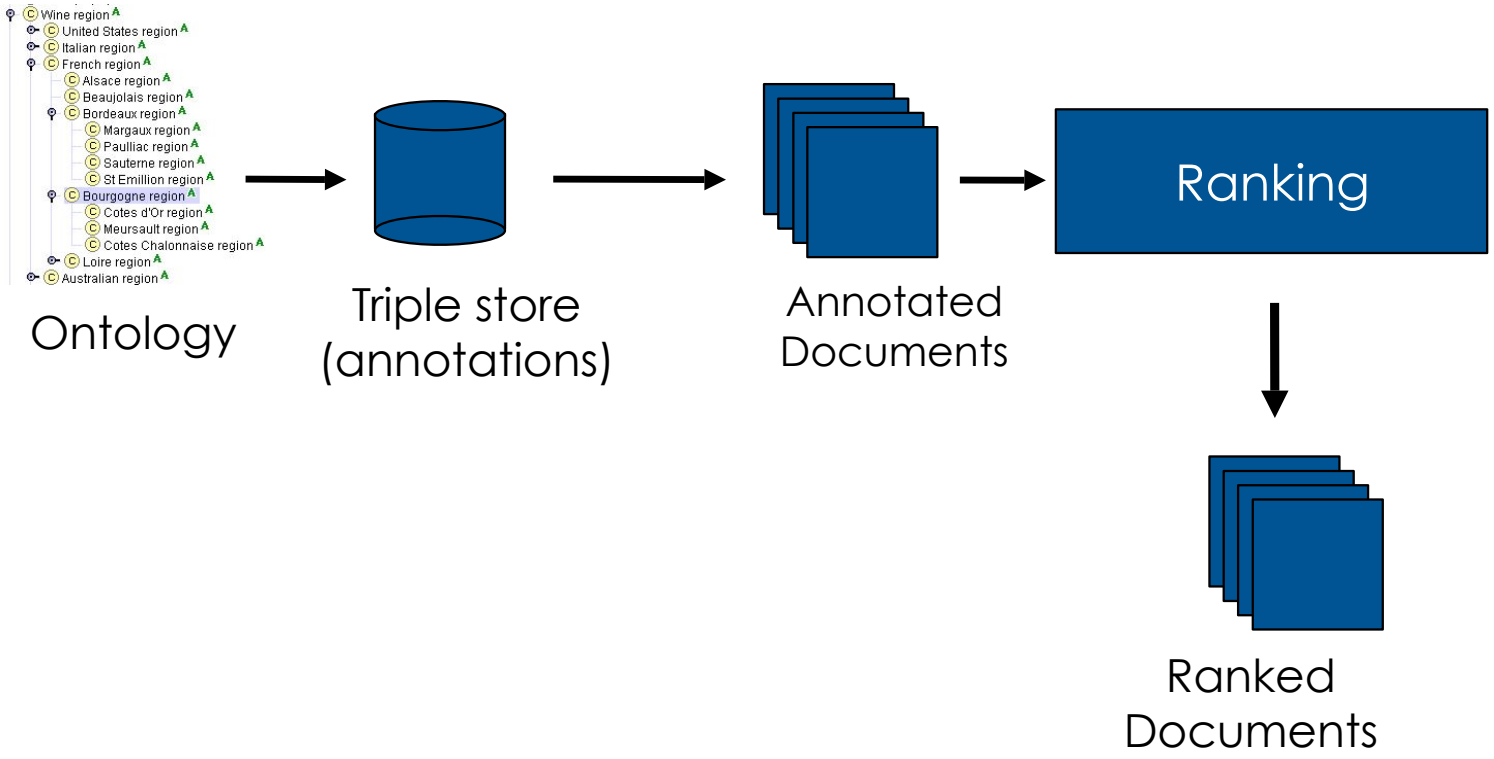
The University Of Sheffield.

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)

ISWC 2007

Searching Documents using ontologies



The University Of Sheffield.

Ontology-based Querying: Issues

- Building an ontology is expensive and needs maintenance
- Metadata generation of documents is
  - Expensive
  - Error prone
- Metadata can cover just part of the material of interest to the users
  - The information not annotated using metadata is irretrievable
    - Often the use people will do of information is impossible to foresee
    - Sometimes Information is impossible to retrieve reliably using automatic methods
- If automatic means are used, often some parts of the knowledge is beyond the current technical capabilities

The University Of Sheffield.

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

The University Of Sheffield.

Results for ontology matching for even reports

- 85% of documents in the first 20 hits a
  - Compare with keywords: 56%
- 40% of relevant documents are in the pages
  - Compare with keywords: 57%

- Ontology matching implies
  - Reading a limited amount of irrelevant dc
  - Risking missing many documents
  - It is possible to *count* the events

The University Of Sheffield.

Issues and Solutions

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

The University Of Sheffield.

Hybrid Search (keywords+ontology)

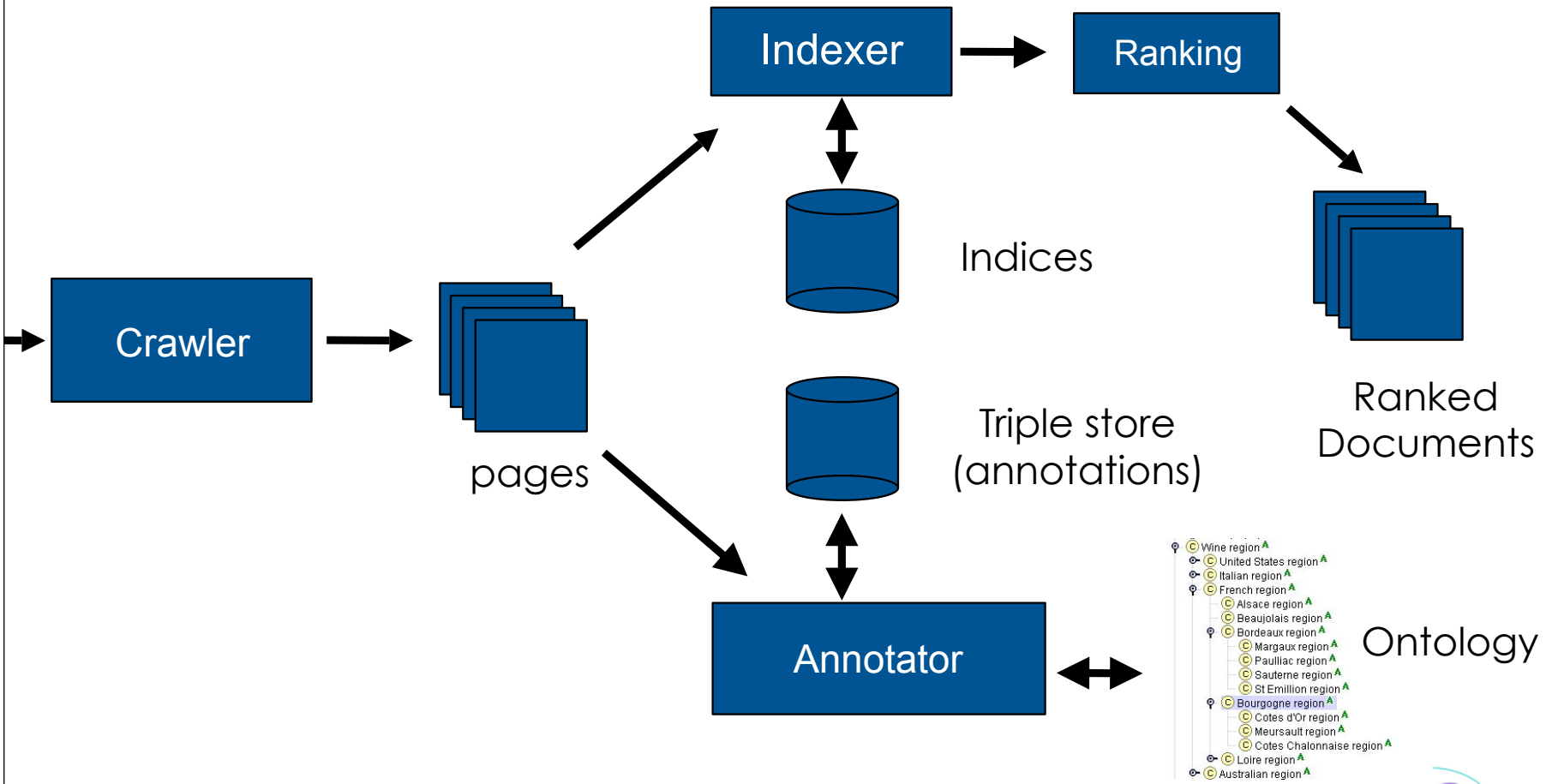
- Mixes keyword and ontology based search
- Ontology based search
- Traditional keyword search
- Keyword in contest of ontology-based annotations
- Potential queries:
  - Return all documents where the word fuel is mentioned
  - Return all documents where the affected part description includes the word fuel
  - Return all documents where the affected part description is similar to "fuel duct"
  - Return all documents where the affected part description is equal to "fuel duct" (URI=XXXXX)

affected parts is concept in ontology

Vitaveska Lanfranchi, Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Daniela Petrelli: Extracting and Searching Knowledge for the Aerospace Industry, in Proc. of 1st European Semantic Technology Conference, Vienna, May 2007

The University Of Sheffield.

Hybrid Indexing/Annotation



Hybrid Search

The University Of Sheffield.

Advantages with Hybrid Search

- Accuracy of Ontology-based searching available
  - When metadata covers information
- Expressiveness of Keyword querying is available
  - For all other cases
- Keyword-in-context available
  - Keyword matching available for matching concepts names
    - e.g. match "fuel" in the description of the removed parts
  - Uses provenance of annotations
    - Portion of document annotated with concepts are stored in 3store
    - Keyword matching applied only on the relevant strings
      - e.g. "fuel" is matched only on snippets of texts annotated as removed parts

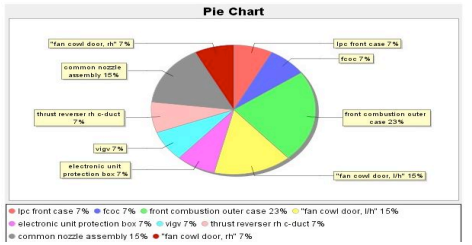
The University Of Sheffield.

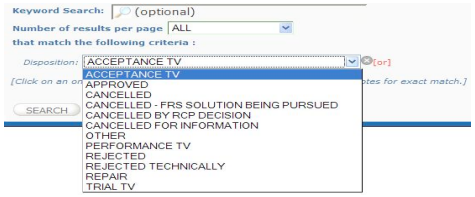
Results for Hybrid Search

- 83% of documents in the first 20 hits are relevant
  - K:56% O:85%
- 85% of relevant documents are in the first 2 pages
  - K: 57% O:47%
- F(1)=84%
  - K:57% O:54%
- Hybrid Search implies
  - Reading a limited amount of irrelevant documents
  - Being able to retrieve easily a very large part of documents

K-Search

- Enables querying documents based on
  - keywords
  - ontology
  - keywords in context
- Enables quantification of unstructured information
- Currently applied to:
  - Event Reports (1998-2004),
  - Technical variances
- Finalist of Rolls-Royce Creativity Award 2007





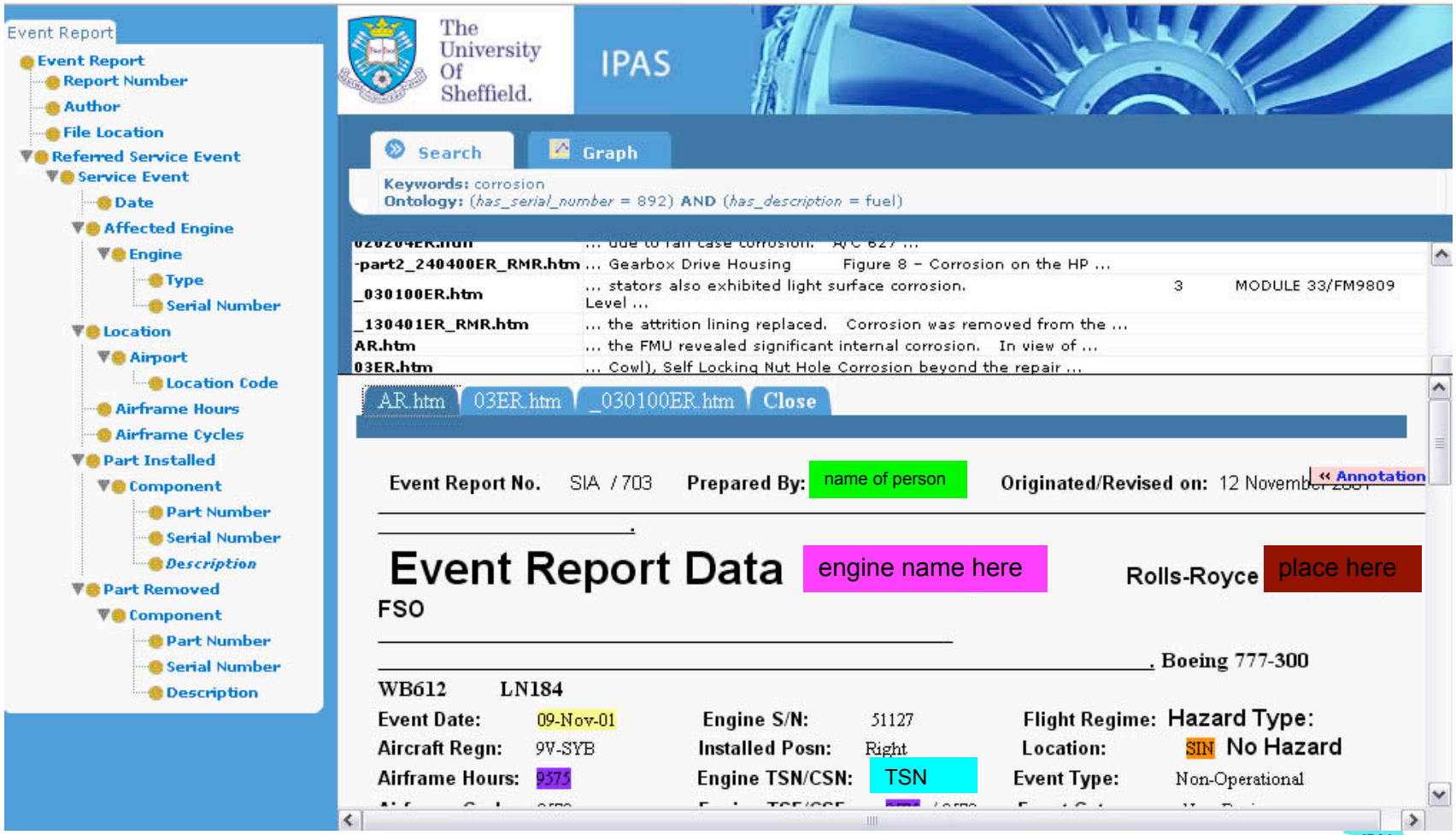
The University Of Sheffield.

Query results

- Results are displayed as a list
- User can click on a document and open it in the lower frame
- The document will be enriched by annotations with attached services
- Multiple documents can be opened in a tab interface

The University Of Sheffield.

Query results



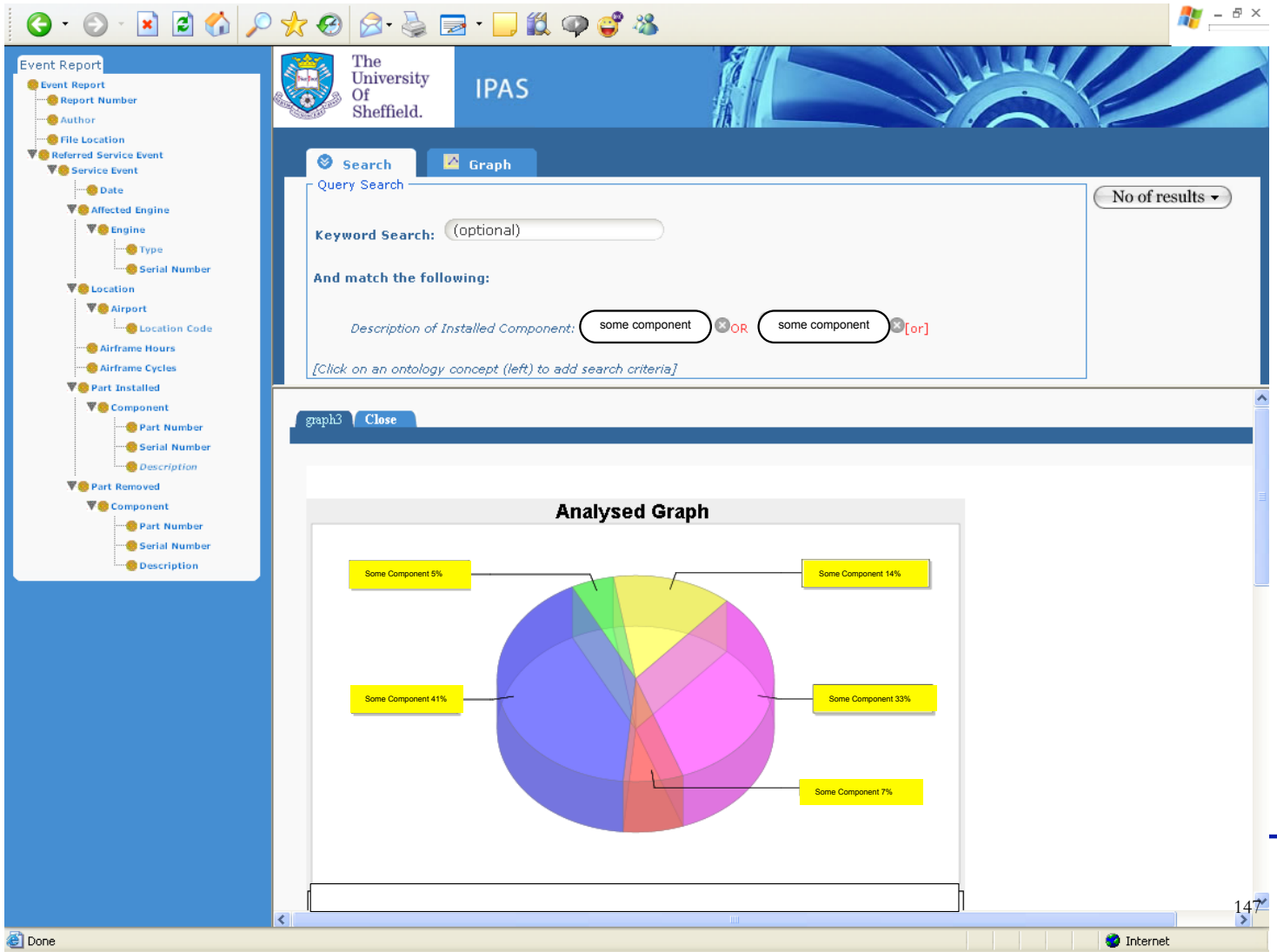
The University Of Sheffield.

Graph visualisation

- Query results can be visualised as a graph
- Users can select graph types
  - Bar chart
  - Pie chart
- Users can select group and subgroup
  - For aggregating results
- Graphs are opened in the tab interface alongside other documents

Graph - example

- Percentage of
- occurrences of installations
- of fuel based parts
- Pie chart



The University Of Sheffield.

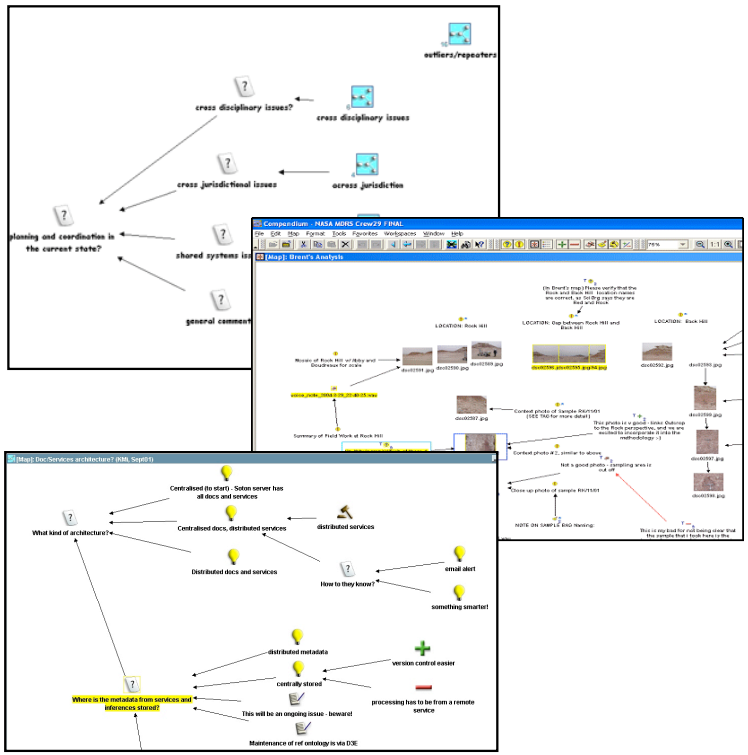
Process Support

- Goals:
  - Supporting users in their tasks
  - Enabling capturing knowledge while knowledge is created

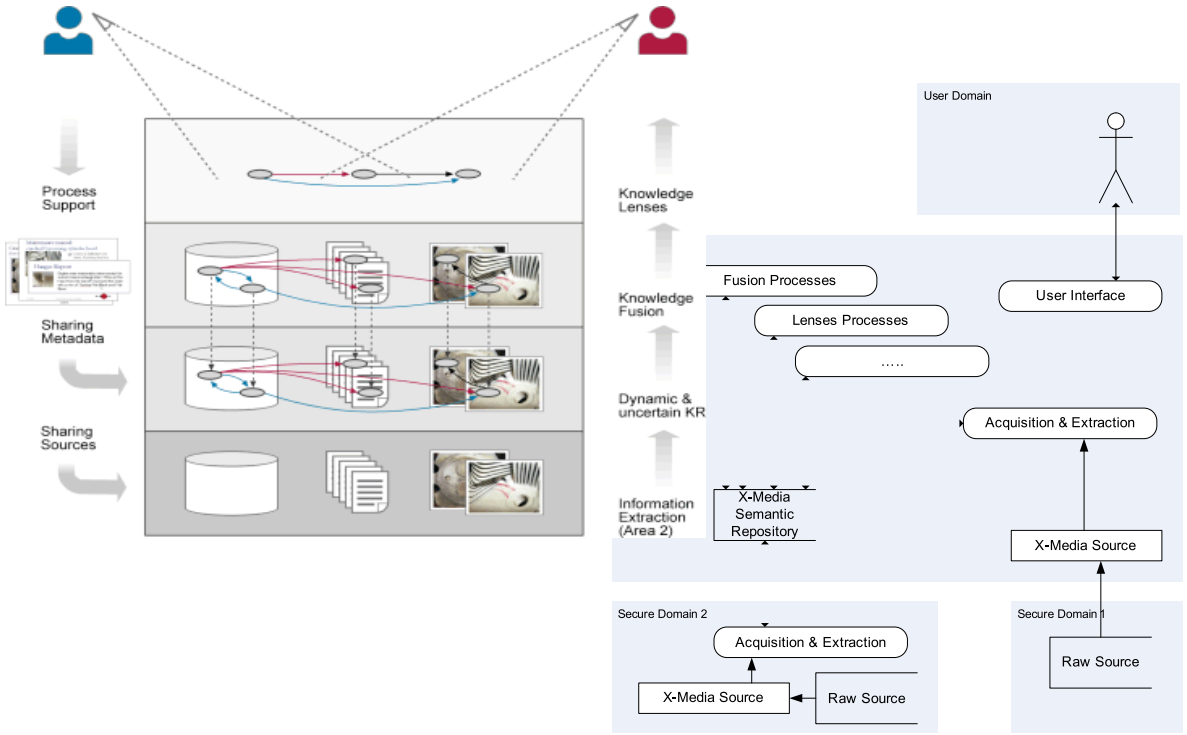
Example: design rationale capturing

- During the discussion, the working group will consider many alternative solutions
  - Those discarded are not in the final document
- When next engine is designed, the group needs to know
  - What solutions were tried (use of titanium)
  - Why they were not adopted (e.g. too high a cost)
  - If the analysis is still true (titanium cost has decreased)

- Compendium (Buckingam-Shum 2002)
- D/Red (Wallace *et al.* 2005, 2007)



ISWC 2007



Other issues: Architectures and privacy

The University Of Sheffield.

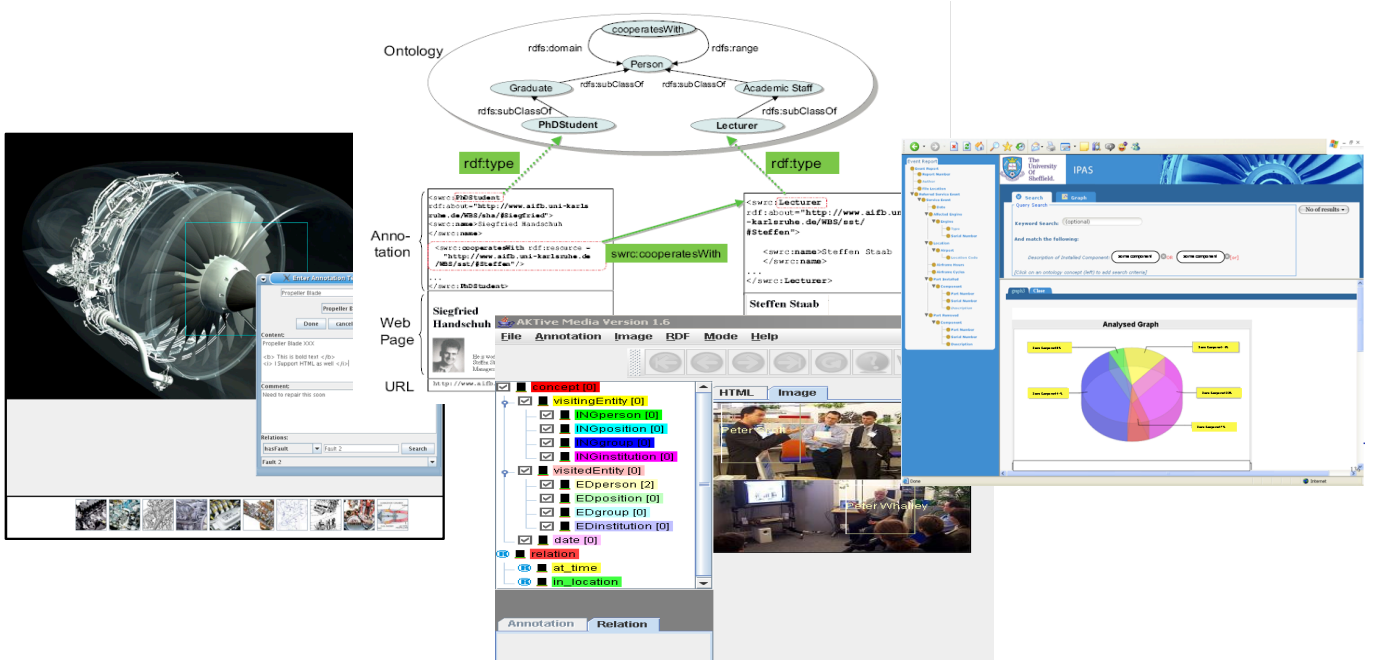
Requirements for KM: Infrastructure

- Infrastructure:
  - Different media cannot easily be shared in the same way as knowledge from one format.
    - A folder of text documents may still be easily sent via email, but a folder of image files may not, and may instead require a shared image repository.
    - For 10 GByte of noise sensor data even an upload to a centralized repository may be out of the question and instead remote access to the underlying data base is to be considered.
  - A complex infrastructure is needed in order to implement knowledge management across media.

The University Of Sheffield.

Privacy and Confidentiality

- Not all information and knowledge is to share
  - Personal knowledge
  - Departmental knowledge
  - Organisational knowledge
  - Public knowledge
- Protecting information is important
- Acquired Knowledge must be shared with care
  - All knowledge must be marked with provenance and confidentiality (Lanfranchi *et al.* 2004)
  - Question: can a piece of knowledge derived also from confidential data (e.g. statistics) be shown to everyone?
    - If not, what do you show to non allowed users? False deductions?



Conclusions

The University Of Sheffield.

Conclusions and Future Work

- Knowledge Management is moving towards large scale
  - Initially expected around 2010 now already happening
- The Semantic WEB offers potentially key technologies to the development of future KM
  - 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!

Future Trends: Web 2.0

- Folksonomies:
  - Easier to use and implement than ontologies
    - SW needs to make the best out of them (Specia *et al.* 2007)
  - Work very well as approximations of ontologies in many applications and tasks
- Blogs
  - The new frontier of knowledge sharing (e.g. Google)
  - Serious risks seen by the companies for information leak and corporate responsibility
    - Whistleblowers and real concerns about putting in writing
  - Semantic blog to acquire and share information

Future Trends: Web 2.0 (ctd)

- Semantic email as a way to trace what is in emails
- Wikis
  - Collaborative working made easier
  - Semantic Wikis a way to acquire and share knowledge in a more effective way
- In general: collaborative thinking of Web 2.0 can potentially impact KM
  - Social aspects:
    - Flink for expert finding
    - Importance of social connection in the current organisation to be enabled, not prevented
  - Semantic Compendium to help capture rationale of design.

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

The University Of Sheffield.

A very Incomplete Bibliography

- 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.
- Vitaveska Lanfranchi, Fabio Ciravegna, Daniela Petrelli: Semantic Web-based Document: Editing and Browsing in AktiveDoc, Proceedings of the 2nd European Semantic Web Conference , Heraklion, Greece, May 29-June 1, 2005
- 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

ISWC 2007

The University Of Sheffield.

A very Incomplete Bibliography (ctd)

- 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.
- C. Goble, S. Bechhofer, L. Carr, D. De Roure, and W. Hall. Conceptual Open Hypermedia = The Semantic Web? In The Second International Workshop on the Semantic Web, pages 44–50, Hong Kong, May 2001
- 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
- Martin Dzbor, John B. Domingue, and Enrico Motta. Magpie - towards a semantic web browser. In Proceedings of the 2nd Intl. Semantic Web Conference, October 2003. Sanibel Island, Forida
- Alexander Maedche, Steffen Staab, Nenad Stojanovic, Rudi Studer, York Sure: SEmantic portAL - The SEAL approach In D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (eds.), Spinning the Semantic Web, pp. 317-359. MIT Press, Cambridge, MA., 2003.

ISWC 2007

The University Of Sheffield.

A very Incomplete Bibliography (ctd)

- Natalya F. Noy and Deborah L. McGuinness: Ontology Development 101: A Guide to Creating Your First Ontology, http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
- Elena Paslaru Bontas, Christoph Tempich, York Sure : OntoCom: A Cost Estimation Model for Ontology Engineering, In: Proceedings of the 5th International Semantic Web Conference (ISWC 2006), November 5-9, 2006, Athens, GA, USA, LNCS. Springer.
- Ajay Chakravarthy, Vita Lanfranchi and Fabio Ciravegna: Cross-media Document Annotation and Enrichment, SAAW2006 - 1st Semantic Authoring and Annotation Workshop, The 5th International Semantic Web Conference (ISWC2006), Athens, GA, USA, Monday, November 6th 200
- R. Gaizauskas and G. Demetriou and P. Artymiuk and P. Willett: Protein Structures and Information Extraction from Biological Texts: The PASTA System, Journal of Bioinformatics 19(1), 135-143, 2003
- Vitaveska Lanfranchi, Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Daniela Petrelli: Extracting and Searching Knowledge for the Aerospace Industry, in Proc. of 1st European Semantic Technology Conference, Vienna, May 2007