Combining Flexible Queries and Knowledge Anchors to facilitate the exploration of Knowledge Graphs

Alex Poulovassilis

Joint work with Riccardo Frosini and Mirko Dimartino
Knowledge Lab, Birkbeck, University of London
Vania Dimitrova and Marwan Al-Tawil
School of Computing, University of Leeds

ICFCA 2017, 13-16 June 2017

Knowledge Lab
Birkbeck UNIVERSITY OF LONDON

Outline of the talk

Introduction and Motivation

Case Study – L4All Ontology and Dataset

Knowledge Anchor Derivation

Extending Flexible Querying with Knowledge Anchors

Conclusions and Future work

Introduction and Motivation

Increasing volumes of graph-structured data arising from many application areas, e.g. RDF linked data

Volumes, complexity and heterogeneity of the data means that users are unlikely to be familiar with its full structure and content

Hence need to be assisted by intelligent tools that support users' interactive exploration

Recent work has proposed flexible querying to find paths through knowledge graphs:

> We have extended SPARQL 1.1 with approximation and relaxation operations – **SPARQL^{AR}** query language

Introduction and Motivation

Features of our flexible querying approach :

- users' queries do not have to match exactly the data structures being queried
- the query system can automatically make changes to a query so as to help the user find relevant information
- query answers are returned in ranked order, in increasing 'distance' from the original query:
    - this is the sum of the costs of all the approximation or relaxation operations that have been applied to the query to obtain this answer
    - the costs of the operations can be application- or user-defined

See Cali et al ODBASE 2014; Frosini et al, Flexible query processing for SPARQL, Semantic Web Journal 8(4) pp 533-563, 2017

Introduction and Motivation

SPARQL^{AR} supports two kinds of flexible querying :

- query relaxation
  - applies a relaxation operation to the query e.g.
    - replacing a class by a superclass
    - replacing a property by a superproperty
  - returns *additional* answers compared to the exact form of the query
- query approximation
  - applies an edit operation to the query e.g. insertion, deletion or substitution of an edge label
  - returns *different* answers compared to the exact form of the query

Introduction and Motivation

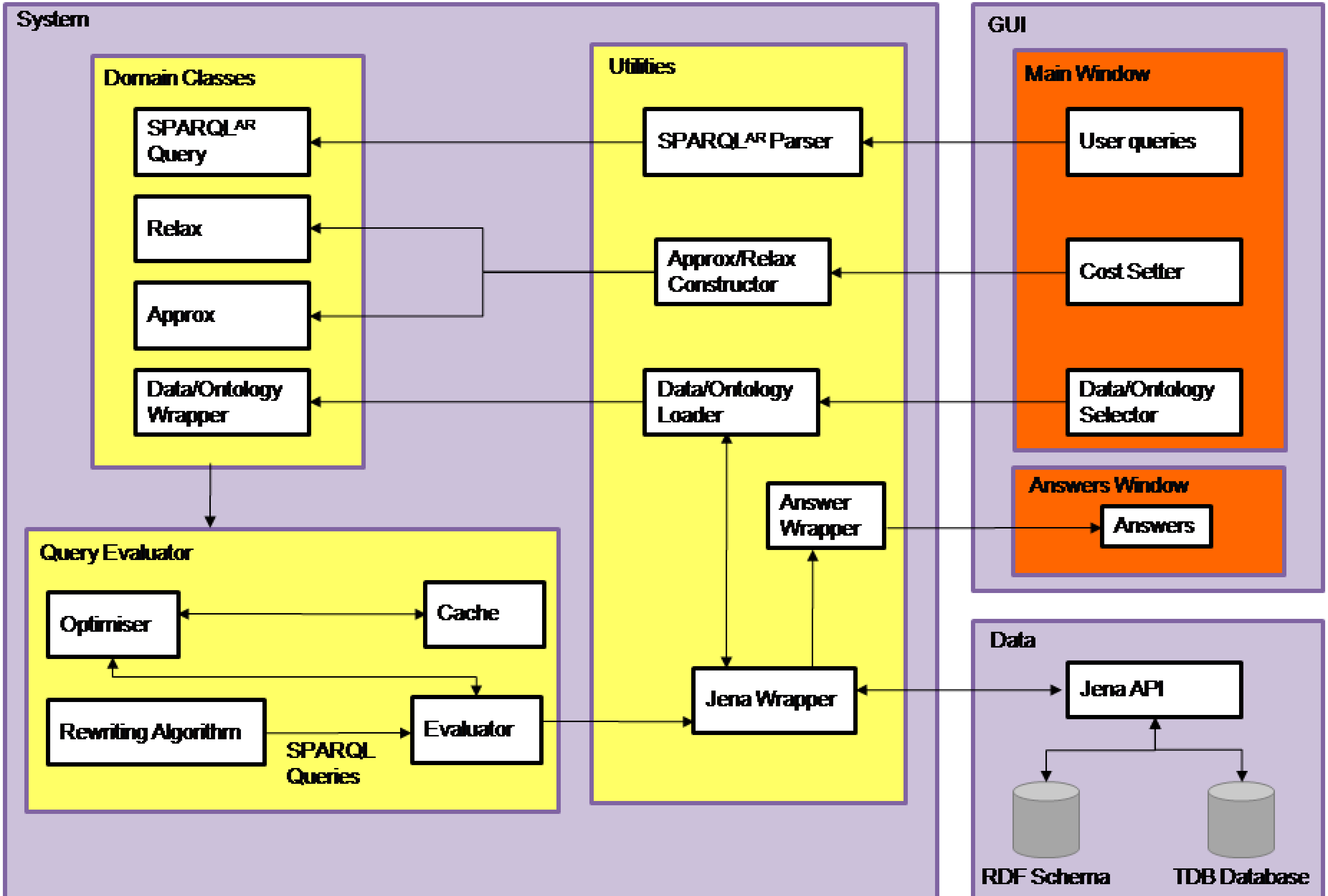
Example SPARQL^{AR} query (see later):

```
SELECT ?WorkEp ?Occ
WHERE {?EdEp rdf:type <http://www.L4All.com/University_Episode>.
        RELAX ( ?EdEp <http://www.L4All.com/qualif>/rdf:type
                <http://www.L4All.com/Information_Systems> ).
        APPROX( ?EdEp <http://www.L4All.com/prereq> ?WorkEp ).
        ?WorkEp  rdf:type  <http://www.L4All.com/Work_Episode>.
        ?WorkEp  <http://www.L4All.com/job>/rdf:type ?Occ}
```

Of course, this is not what an end-user would enter, but it could be the SPARQL^{AR} query generated from the user's interaction with a graphical user interface

See e.g. Poulovassilis et al, Flexible Querying of Lifelong Learner Metadata, IEEE Trans. Learning Technologies, 5(2) pp 117-129, 2012

SPARQL^{AR} system architecture



Introduction and Motivation

Supporting such flexible query processing over knowledge graphs brings several benefits:

- automatic correction of users' erroneous queries
- finding additional relevant answers that the user may be unaware of, due to data modelling heterogeneities
- generating new queries which may return unexpected results and bring new insights

Introduction and Motivation

Although flexible query processing allows broadening a user’s perspective of the knowledge domain, it can return a large number of results all at the same 'distance' from the user’s original query

Therefore, a key challenge is how to *facilitate users' meaning making from flexible query results*

This requires supporting users’ knowledge expansion starting from entities that are close to the users’ cognitive structures

Introduction and Motivation

Detecting which entities a user may be familiar with (e.g. by analysing interaction logs) is a computationally intensive task

When the user has had limited interaction with the system, there is also the 'cold start' problem

Other ways are needed for automatically identifying which entities may be familiar for the user and hence may be good *Knowledge Anchors* for information exploration in a knowledge graph

Introduction and Motivation

Recent work has proposed an approach that adopts the Cognitive Science notion of *basic-level objects* in domain taxonomies:

> See M. Al-Tawil, V. Dimitrova, D. Thakker, and B. Bennett, Identifying knowledge anchors in a data graph, ACM Conf. on Hypertext and Social Media, 2016

That work has developed a formal framework for identifying knowledge anchors (KAs) in knowledge graphs, applying two complementary categories of metrics:

> *distinctiveness* and *homogeneity*

Identifying Knowledge Anchors (HT2016)

UNIVERSITY OF LEEDS

Distinctiveness

Input: $DG = \langle V, E, P \rangle, e \in E$

for all $v \in \{C\}$ **do**

$V' :=$ the set of all $v' : v' \subseteq v$

for all $v'_e : \exists \langle v'_e, e, v' \rangle$ **do**

$N_e :=$ set of all $\langle v'_e, e, v' \rangle : v' \in V'$

$M_e :=$ set of all $\langle v'_e, e, v_a \rangle : v_a \in V$

$AV_{v'_e} := |N_e| / |M_e|$

$CAC_{v'_e} := (|N_e| / |M_e|) \cdot (|N_e| / |V'|)$

$CU_{v'_e} := (|N_e| / |V'|)^2 - (|M_e| / |V|)^2$

$AV_v := AV_v + AV_{v'_e}$

$CAC_v := CAC_v + CAC_{v'_e}$

$CU_v := CU_v + CU_{v'_e}$

end for

end for

Output: AV_v, CAC_v, CU_v for all $v \in \{T \cup C\}$

Homogeneity

Input: $DG = \langle V, E, P \rangle, e \in E$

for all $v \in \{C\}$ **do**

$V' :=$ the set of all $v' : v' \subseteq v$

for all $(v', v'') : v' \in V' \wedge v'' \in V'$ **do**

$V'_e := \{v'_e : \exists \langle v'_e, e, v' \rangle\}$

$V''_e := \{v''_e : \exists \langle v''_e, e, v'' \rangle\}$

$I := V'_e \cap V''_e$

$U := V'_e \cup V''_e$

$CN_{v',v''} := |I|$

$Jac_{v',v''} := |I| / |U|$

$Cos_{v',v''} := |I| / (\sqrt{|V'_e|} \cdot \sqrt{|V''_e|})$

$CN_v = CN_v + CN_{v',v''}$

$Jac_v = Jac_v + Jac_{v',v''}$

$Cos_v = Cos_v + Cos_{v',v''}$

end for

$CN_v = CN_v / (|V'|.(|V'|-1)/2)$

$Jac_v = Jac_v / (|V'|.(|V'|-1)/2)$

$Cos_v = Cos_v / (|V'|.(|V'|-1)/2)$

end for

Output: CN_v, Jac_v, Cos_v for all $v \in \{C\}$

Introduction and Motivation

We draw on these two strands of work:

- flexible querying of graph-structured data
- identification of KAs for graph exploration

to support users in incrementally querying, exploring and learning from large, complex knowledge graphs

We illustrate this integrative approach through a case study in exploring Learning and Career Options

For details see: Poulovassilis et al, Combining Flexible Queries and Knowledge Anchors to facilitate the exploration of Knowledge Graphs, Proceedings 5th International Workshop on Intelligent Exploration of Semantic Data (IESD), at ISWC 2016

Case Study – L4All Ontology and Dataset

The L4All project aimed to provide learners with access to information and resources that would support them in exploring learning and career opportunities and in planning and reflecting on their learning

It brought together experts from lifelong learning and careers guidance, content providers, and groups of students and tutors

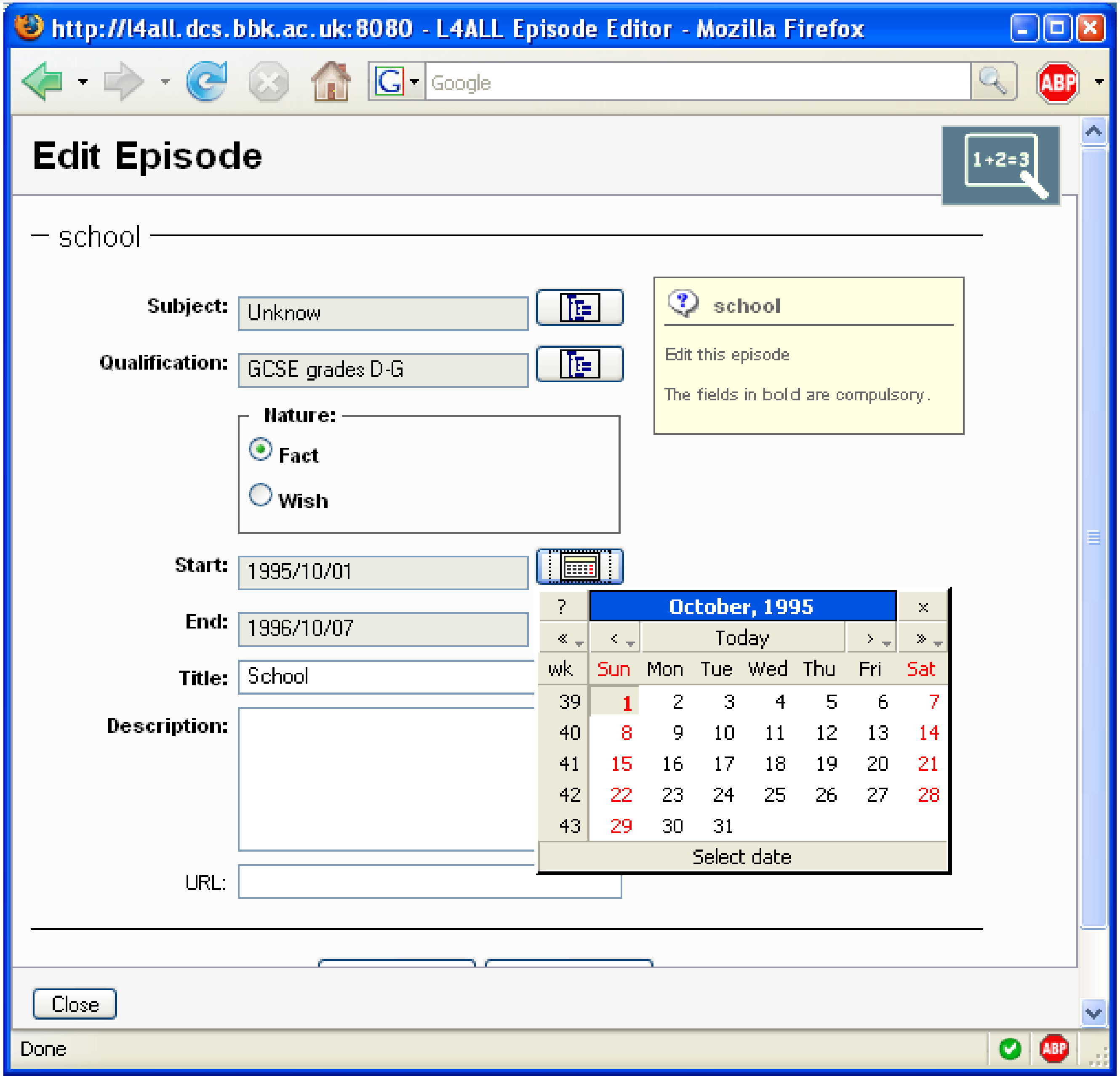
The L4All pilot system allowed users to record their learning, work and life experiences within a ‘timeline’

Users’ timelines are encoded in RDF/S. Educational and Work episodes need to be annotated by the user with a primary and possibly a secondary classification. These classifications are drawn from standard taxonomies of the UK Office for National Statistics

- Subject
  - Architecture,_Building_and_Planning
  - Biological_Sciences
  - Business_and_Administrative_Studies
  - Creative_Arts_and_Design
  - East,_Asiatic,_African,_American_and_Australian_Languages,_
  - Education
  - Engineering
  - European_Language,_Literature_and_related_subjects
  - Historical_and_Philosophical_Studies
  - Law
  - Linguistics,_Classics_and_related_subjects
  - Mass_Communications_and_Documentation
  - Mathematical_and_Computer_Sciences
    - Artificial_Intelligence
    - Computer_Science
    - Information_Systems
    - Mathematics
    - Operational_Research
    - Other_in_Mathematics_and_Computer_Sciences
    - Software_Engineering
    - Statistics
  - Medicine_and_Dentistry
  - Physical_Sciences
  - Social_Studies
  - Subjects_allied_to_Medicine


Occupation
Administrative_and_Secretarial_Occupations
Associate_Professional_and_Technical_Occupations
Elementary_Occupations
Managers_and_Senior_Officials
Personal_Service_Occupations
Caring_Personal_Service_Occupations
Leisure_and_Other_Personal_Service_Occupations
Process,_Plant_and_Machine_Operatives
Professional_Occupations
Business_and_Public_Service_Professionals
Architects,_Town_Planners_and_Surveyors
Business_and_Statistical_Professionals
Legal_Professionals
Librarians_and_Related_Professionals
Public_Service_Professionals
Health_Professionals
Science_and_Technology_Professionals
Engineering_Professionals
Chemical_Engineers
Civil_Engineers
Design_and_Development_Engineers
Electrical_Engineers
Electronics_Engineers
Engineering_Professionals_N.E.C.
Mechanical_Engineers
Planning_and_Quality_Control_Engineers
Production_and_Process_Engineers
Information_and_Communication_Technology_Professionals
IT_Strategy_and_Planning_Professionals
Software_Professionals
Science_Professionals
Teaching_and_Research_Professionals
Research_Professionals
Teaching_Professionals
Sales_and_Customer_Service_Occupations
Skilled_Trades_Occupations

Entering/Editing Episodes



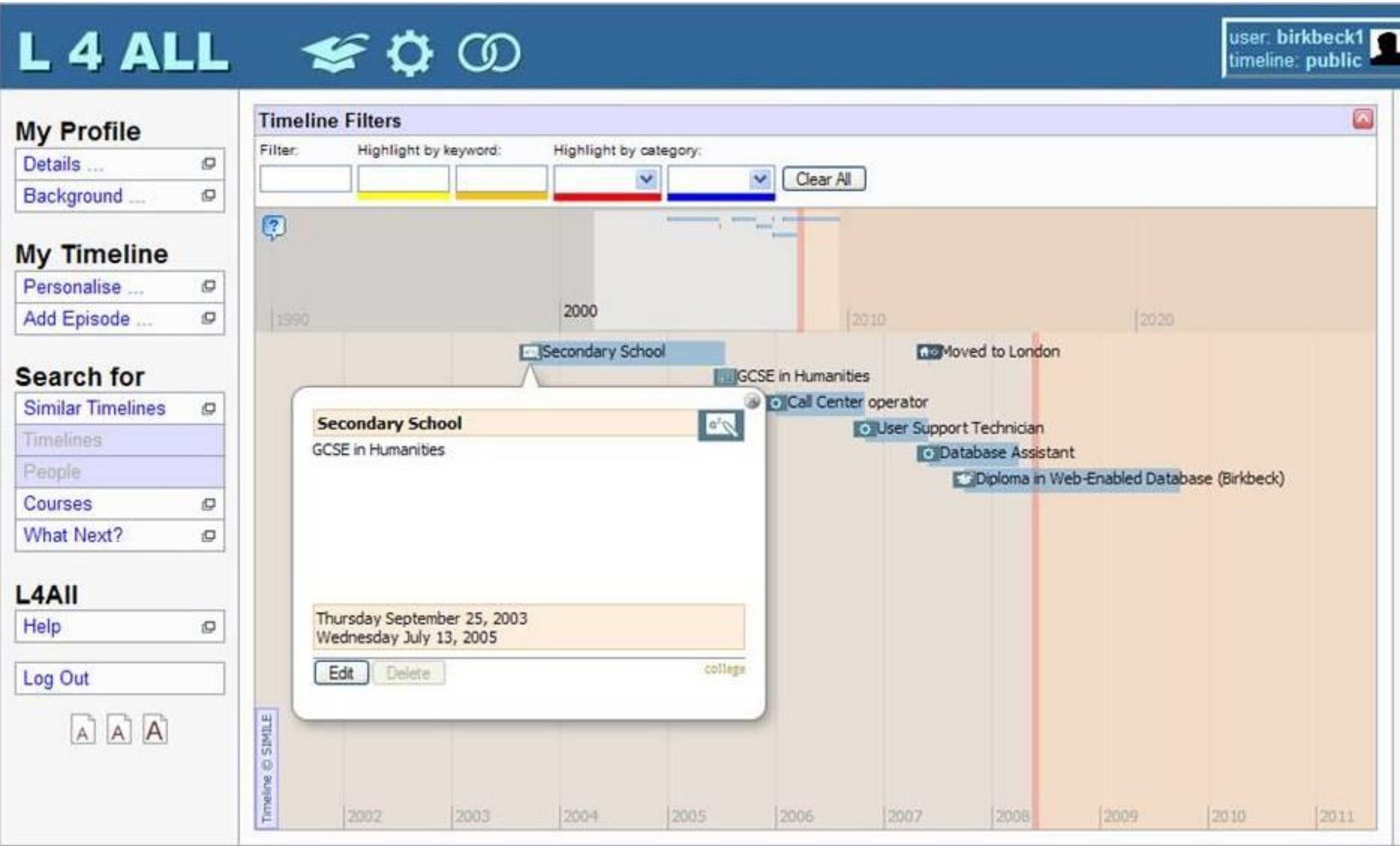
Entering/Editing Episodes

A key finding from the L4All user studies was that participants experienced problems in classifying their episodes

- they had particular difficulty in finding the relevant element within a taxonomy
- they spent a significant amount of time trying to specify the appropriate – to the best of their knowledge – classification for their episodes

Adding to the difficulty was the fact that the need for supplying these classifications were not immediately apparent – it only become apparent when they later started using the search facilities

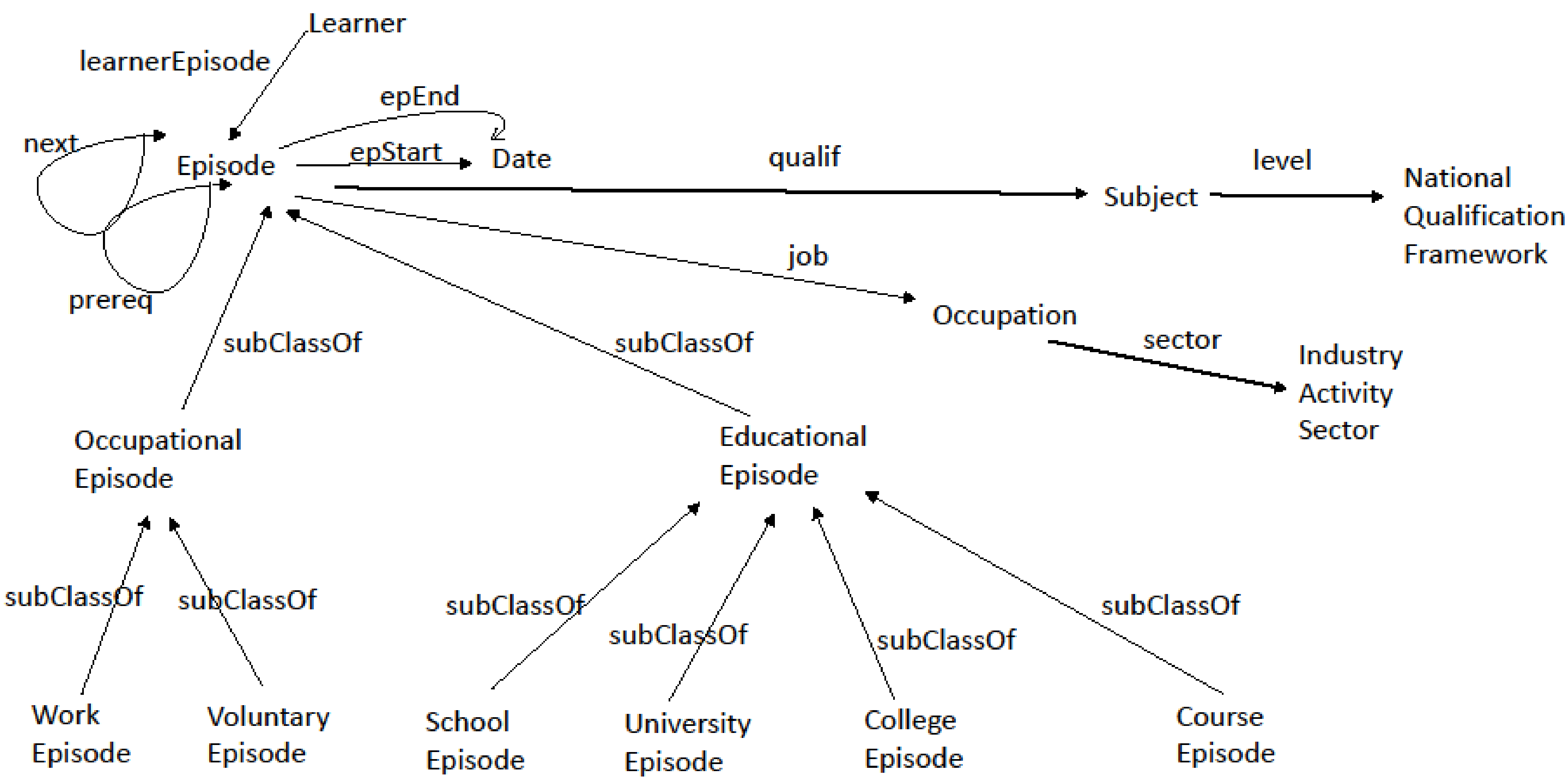
Main UI screen – showing a whole timeline



Case Study – L4All Ontology and Dataset

- Users are able to search over the timelines of other learners and professionals (if a timeline has been made public by its owner)
  - gives a repertoire of learning and work possibilities that may not otherwise have been considered
  - presents successful learners as role models to inspire confidence and a sense of opportunity
- In the original pilot system, *similarity measures* were used for comparing users' timelines against the user's search criteria
- This had a number of drawbacks, largely arising from the rigidity of the similarity matching algorithms employed and users' lack of understanding of how they worked
- Led to further work that investigated the use of *flexible query processing* techniques based on query approximation and relaxation to support users' search over the timeline data
  - See Poulovassilis et al IEEE Trans. Learning Technologies, 5(2), 2012

Case Study – L4All Ontology and Dataset



Case Study – L4All Ontology and Dataset

Class hierarchy	Depth	No. of classes
Episode	3	9
Subject	3	161
Occupation	4	465
National Qualification Framework	3	36
Industry Activity Sector	3	22

Knowledge Anchor Derivation

We generated the KAs from the L4All ontology combined with an RDF dataset comprising data relating to 1700 timelines

KAs aim to represent *familiar* and *highly inclusive entities* in the graph from which links to new knowledge can be made

This new knowledge can take meaning by becoming linked to existing concepts within the user's cognitive structures

We considered two types of relationships:

- *hierarchical relationships* denoting membership between the subject and object of an RDF triple (rdfs:subClassOf, rdf:type)
- *domain-specific relationships*, which are properties other than the hierarchical relationships

Knowledge Anchor Derivation

We adopted two complementary groups of metrics to identify KAs (see Al-Tawil et al, ACM Conf. on Hypertext and Social Media, 2016 for a detailed description of the metrics and algorithms):

Distinctiveness metrics identify the most differentiated categories, whose attributes are associated with the category members but not with members of other categories. We use three distinctiveness metrics: Attribute Validity (AV), Category-Attribute Collocation (CAC), Category Utility (CU).

Homogeneity metrics identify categories whose members share many attributes. We use three set-based similarity metrics: Common Neighbors (CN), Jaccard (Jac), Cosine (Cos).

Identifying Knowledge Anchors (HT2016)

UNIVERSITY OF LEEDS

Distinctiveness

Input: $DG = \langle V, E, P\rangle, e \in E$

for all $v \in \{C\}$ **do**

$V' :=$ the set of all $v' : v' \subseteq v$

for all $v'_e : \exists\langle v'_e, e, v'\rangle$ **do**

$N_e :=$ set of all $\langle v'_e, e, v'\rangle : v' \in V'$

$M_e :=$ set of all $\langle v'_e, e, v_a\rangle : v_a \in V$

$AV_{v'_e} := |N_e| / |M_e|$

$CAC_{v'_e} := (|N_e| / |M_e|) \cdot (|N_e| / |V'|)$

$CU_{v'_e} := (|N_e| / |V'|)^2 - (|M_e| / |V|)^2$

$AV_v := AV_v + AV_{v'_e}$

$CAC_v := CAC_v + CAC_{v'_e}$

$CU_v := CU_v + CU_{v'_e}$

end for

end for

Output: AV_v, CAC_v, CU_v for all $v \in \{T \cup C\}$

Homogeneity

Input: $DG = \langle V, E, P\rangle, e \in E$

for all $v \in \{C\}$ **do**

$V' :=$ the set of all $v' : v' \subseteq v$

for all $(v', v'') : v' \in V' \wedge v'' \in V'$ **do**

$V'_e := \{v'_e : \exists\langle v'_e, e, v'\rangle\}$

$V''_e := \{v''_e : \exists\langle v''_e, e, v''\rangle\}$

$I := V'_e \cap V''_e$

$U := V'_e \cup V''_e$

$CN_{v',v''} := |I|$

$Jac_{v',v''} := |I| / |U|$

$Cos_{v',v''} := |I| / (\sqrt{|V'_e|} \cdot \sqrt{|V''_e|})$

$CN_v = CN_v + CN_{v',v''}$

$Jac_v = Jac_v + Jac_{v',v''}$

$Cos_v = Cos_v + Cos_{v',v''}$

end for

$CN_v = CN_v / (|V'|.(|V'|-1)/2)$

$Jac_v = Jac_v / (|V'|.(|V'|-1)/2)$

$Cos_v = Cos_v / (|V'|.(|V'|-1)/2)$

end for

Output: CN_v, Jac_v, Cos_v for all $v \in \{C\}$

Knowledge Anchor Derivation

- The six metrics are calculated for each class in the graph, considering both its hierarchical and its domain-specific relationships
- Hence, for each class we obtain 12 scores that rate that entity's suitability as a KA
- We selected entities with at least 50% non-zero scores, subject to the constraint that a KA should have at least one non-zero score from the subset of hierarchical relationships and at least one non-zero score from the subset of domain-specific relationships

Knowledge Anchor Derivation

- For example, the KAs identified within the Subject hierarchy include:

  Architecture,_Building_and_Planning
  Biological_Sciences
  Business_and_Administrative_Studies
  Creative_Arts_and_Design
  Education
  European_Language,_Literature_and_related_subjects
  Lingustics,_Classics_and_related_subjects
  Mathematical_and_Computer_Sciences

Knowledge Anchor Derivation

- The KAs identified within the Occupation hierarchy include:

  Administrative_Occupations
  Associate_Professional_and_Technical_Occupations
  Corporate_Managers
  Managers_and_Senior_Officials
  Personal_Service_Occupations
  Professional_Occupations
  Science_and_Technology_Professionals
  Teaching_and_Research_Professionals

Extending Flexible Querying with KAs

- Suppose the user is currently studying for a Foundation Degree in IT and wishes to find out what possible future job choices there are by seeing what other people with qualifications in Information Systems, or similar, have gone on to do
- This can be undertaken by evaluating the following SPARQL^{AR} query
- Before running the query, the user elects (through the system's GUI) to apply two of the edit operations available as part of the APPROX operator:
  - Insertion of an edge label; Substitution of an edge label;
- and also selects one relaxation operation from those available as part of the RELAX operator:
  - Replacement of a subclass by its immediate superclass;
- All edit/relaxation operations are set to have a cost of 1

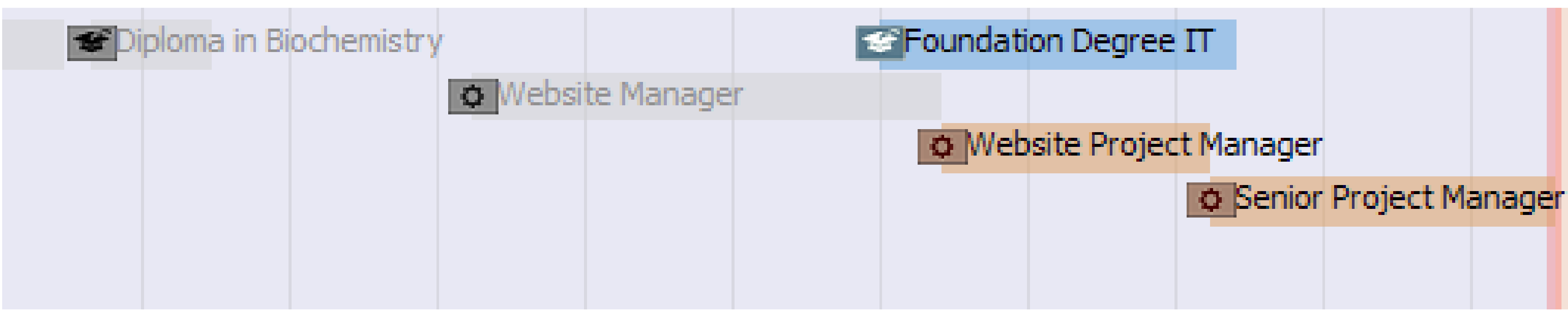
Extending Flexible Querying with KAs

```
SELECT ?WorkEp ?Occ
WHERE {?EdEp  rdf:type  <http://www.L4All.com/University_Episode>.
      RELAX ( ?EdEp <http://www.L4All.com/qualif>/rdf:type
                  <http://www.L4All.com/Information_Systems> ).
      APPROX( ?EdEp <http://www.L4All.com/prereq> ?WorkEp ).
      ?WorkEp  rdf:type  <http://www.L4All.com/Work_Episode>.
      ?WorkEp  <http://www.L4All.com/job>/rdf:type ?Occ}
```

First 20 exact answers

A_4_E_14_22	Research_Professionals
A_1_E_8_98	IT_User_Support_Technicians
A_3_E_5_37	Software_Professionals
A_7_E_7_92	Engineering_Technicians
A_7_E_7_4	Quality_Assurance_Technicians
A_7_E_7_60	Quality_Assurance_Technicians
A_2_E_6_11	Purchasing_Managers
A_8_E_5_88	Pensions_and_Insurance_Clerks
A_8_E_6_88	Physicists,_Geologists_and_Meterologists
A_1_E_8_58	IT_User_Support_Technicians
A_4_E_14_14	Registrars_and_Senior_Administrators_of_Educational_Establishments
A_7_E_7_52	Architectural_Technologists_and_Town_Planning_Technicians
A_4_E_14_78	Primary_and_Nursery_Education_Teaching_Professionals
A_7_E_7_20	Laboratory_Technicians
A_4_E_14_46	Scientific_Researchers
A_7_E_7_84	IT_Service_Delivery_Occupations
A_8_E_5_24	Stock_Control_Clerks
A_8_E_6_24	Electronics_Engineers
A_3_E_5_21	Software_Professionals
A_2_E_6_3	Advertising_and_Public_Relations_Managers

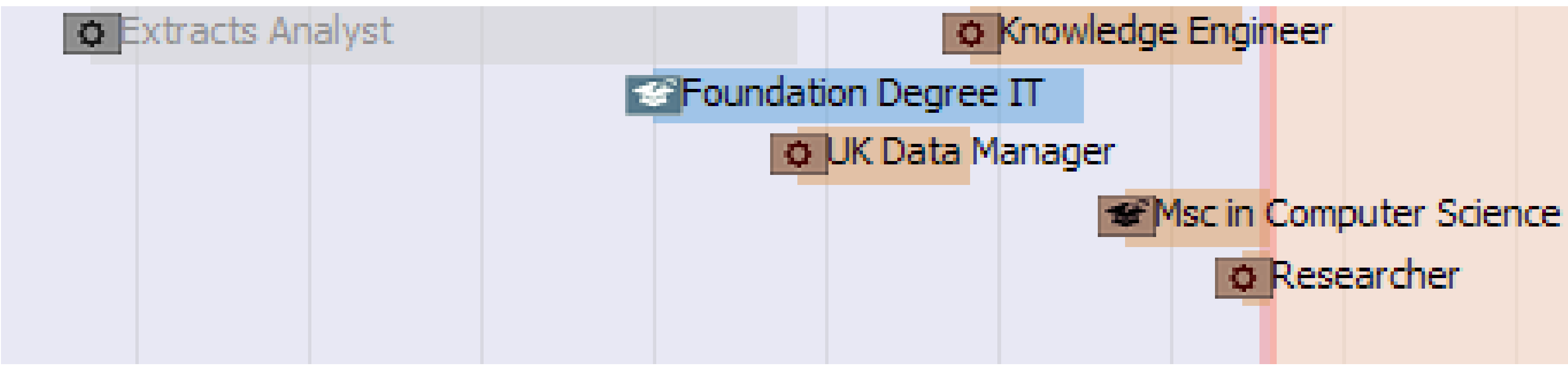
Clicking on an episode URI (in the first column of the results set) the can user explore a whole timeline



First 20 answers at distance 1

A_2_E_6_51	Personnel,_Training_and_Industrial_Relations_Managers	1
A_7_E_7_12	IT_User_Support_Technicians	1
A_8_E_5_56	Market_Research_Interviewers	1
A_7_E_7_44	Science_and_Engineering_Technicians	1
A_8_E_6_72	Civil_Engineers	1
A_8_E_5_64	Library_Assistants/Clerks	1
A_8_E_5_16	Transport_and_Distribution_Clerks	1
A_7_E_7_28	IT_Service_Delivery_Occupations	1
A_8_E_6_8	Mechanical_Engineers	1
A_8_E_5_96	Filing_and_Other_Records_Assistants/Clerks	1
A_2_E_6_35	Research_and_Development_Managers	1
A_8_E_6_40	Planning_and_Quality_Control_Engineers	1
A_4_E_14_6	Social_Science_Researchers	1
A_8_E_5_8	Library_Assistants/Clerks	1
A_1_E_8_50	IT_User_Support_Technicians	1
A_7_E_7_68	IT_User_Support_Technicians	1
A_1_E_8_74	IT_User_Support_Technicians	1
A_1_E_8_34	IT_User_Support_Technicians	1
A_7_E_7_60	Quality_Assurance_Technicians	1
A_3_E_5_69	Software_Professionals	1

Clicking on an episode URI (in the first column of the results set) the user can explore a whole timeline



First 20 answers at distance 2

A_2_E_6_51	Personnel,_Training_and_Industrial_Relations_Managers	2
A_7_E_7_12	IT_User_Support_Technicians	2
A_8_E_5_56	Market_Research_Interviewers	2
A_7_E_7_44	Science_and_Engineering_Technicians	2
A_8_E_6_72	Civil_Engineers	2
A_8_E_5_64	Library_Assistants/Clerks	2
A_8_E_5_16	Transport_and_Distribution_Clerks	2
A_7_E_7_28	IT_Service_Delivery_Occupations	2
A_8_E_6_8	Mechanical_Engineers	2
A_8_E_5_96	Filing_and_Other_Records_Assistants/Clerks	2
A_2_E_6_35	Research_and_Development_Managers	2
A_8_E_6_40	Planning_and_Quality_Control_Engineers	2
A_4_E_14_6	Social_Science_Researchers	2
A_8_E_5_8	Library_Assistants/Clerks	2
A_1_E_8_50	IT_User_Support_Technicians	2
A_7_E_7_68	IT_User_Support_Technicians	2
A_1_E_8_74	IT_User_Support_Technicians	2
A_1_E_8_34	IT_User_Support_Technicians	2
A_7_E_7_60	Quality_Assurance_Technicians	2
A_3_E_5_69	Software_Professionals	2

Extending Flexible Querying with KAs

- It is evident that, although it can return relevant and useful answers for the user, this kind of incremental flexible querying can easily result in information overload.
- Moreover, the user may be unfamiliar with some of the specialist terminology relating to occupations.
- The user will also gain little insight into the relationships between the different occupations being suggested and how they are categorised within the broader context of the Occupation hierarchy

- Repeating the above query and user interactions, we consider an alternative presentation of the results as paths within the Occupation hierachy, rooted at the nearest Knowledge Anchor

First 20 answers at dist. 0, under nearest KA

IT_Service_Delivery_Occupations A_7_E_7_84 (0)
    IT_User_Support_Technicians A_1_E_8_98, A_1_E_8_58 (0)
Science_and_Engineering_Technicians
    Engineering_Technicians A_7_E_7_92 (0)
    Quality_Assurance_Technicians A_7_E_7_4, A_7_E_7_60 (0)
    Laboratory_Technicians A_7_E_7_20 (0)
Draughtspersons_and_Building_Inspectors
    Architectural_Technologists_and_Town_Planning_Technicians A_7_E_7_52 (0)
Corporate_Managers
    Functional_Managers
        Advertising_and_Public_Relations_Managers A_2_E_6_3 (0)
        Purchasing_Managers A_2_E_6_11 (0)
Information_and_Communication_Technology_Professionals
    Software_Professionals A_3_E_5_37, A_3_E_5_21 (0)

Cont'd

Science_Professionals

- Physicists,_Geologists_and_Meterologists A_8_E_6_88 (0)

Engineering_Professionals

- Electronics_Engineers A_8_E_6_24 (0)

Research_Professionals A_4_E_14_22 (0)

- Scientific_Researchers A_4_E_14_46 (0)

Teaching_Professionals

- Registrars_and_Senior_Administrators_of_Educational_Establishments A_4_E_14_14 (0)
- Primary_and_Nursery_Education_Teaching_Professionals A_4_E_14_78 (0)

Administrative_Occupations:_Records

- Pensions_and_Insurance_Clerks A_8_E_5_88 (0)
- Stock_Control_Clerks A_8_E_5_24 (0)

First 20 answers at dist. 0 & 1, under nearest KA

IT_Service_Delivery_Occupations A_7_E_7_84 (0), **A_7_E_7_28 (1)**

- IT_User_Support_Technicians A_1_E_8_98, A_1_E_8_58 (0), **A_7_E_7_12, A_1_E_8_50, A_7_E_7_68, A_1_E_8_74, A_1_E_8_34 (1)**

Science_and_Engineering_Technicians **A_7_E_7_44 (1)**

- Engineering_Technicians A_7_E_7_92 (0)
- Quality_Assurance_Technicians A_7_E_7_4, A_7_E_7_60 (0), **A_7_E_7_60 (1)**
- Laboratory_Technicians A_7_E_7_20 (0)

Draughtspersons_and_Building_Inspectors

- Architectural_Technologists_and_Town_Planning_Technicians A_7_E_7_52 (0)

Corporate_Managers

- Functional_Managers
  - Advertising_and_Public_Relations_Managers A_2_E_6_3 (0)
  - Purchasing_Managers A_2_E_6_11 (0)
  - **Personnel,_Training_and_Industrial_Relations_Managers A_2_E_6_51 (1)**
  - **Research_and_Development_Managers A_2_E_6_35 (1)**

Information_and_Communication_Technology_Professionals

- Software_Professionals A_3_E_5_37, A_3_E_5_21 (0), **A_3_E_5_69, A_8_E_5_96 (1)**

Cont’d

Science_Professionals
- Physicists,_Geologists_and_Meterologists A_8_E_6_88 (0)

Engineering_Professionals
- Electronics_Engineers A_8_E_6_24 (0)
- **Civil_Engineers A_8_E_6_72 (1)**
- **Mechanical_Engineers A_8_E_6_8 (1)**
- **Planning_and_Quality_Control_Engineers A_8_E_6_40 (1)**

Research_Professionals A_4_E_14_22 (0)
- Scientific_Researchers A_4_E_14_46 (0)

Teaching_Professionals
- Registrars_and_Senior_Administrators_of_Educational_Establishments A_4_E_14_14 (0)
- Primary_and_Nursery_Education_Teaching_Professionals A_4_E_14_78 (0)

Administrative_Occupations:_Records
- Pensions_and_Insurance_Clerks A_8_E_5_88 (0)
- Stock_Control_Clerks A_8_E_5_24 (0)
- **Market_Research_Interviewers A_8_E_5_56 (1)**
- **Library_Assistants/Clerks A_8_E_5_64 (1)**
- **Transport_and_Distribution_Clerks A_8_E_5_16 (1)**
- **Filing_and_Other_Records_Assistants/Clerks A_8_E_5_96 (1)**

First 20 answers at dist. 0, under all KAs

Associate_Professional_and_Technical_Occupations
- Science_and_Technology_Associate_Professional
  - IT_Service_Delivery_Occupations A_7_E_7_84 (0)
    - IT_User_Support_Technicians A_1_E_8_98, A_1_E_8_58 (0)
  - Science_and_Engineering_Technicians
    - Engineering_Technicians A_7_E_7_92 (0)
    - Quality_Assurance_Technicians A_7_E_7_4, A_7_E_7_60 (0)
    - Laboratory_Technicians A_7_E_7_20 (0)
- Draughtspersons_and_Building_Inspectors
  - Architectural_Technologists_and_Town_Planning_Technicians A_7_E_7_52 (0)

Managers_and_Senior_Officials
- Corporate_Managers
  - Functional_Managers
    - Advertising_and_Public_Relations_Managers A_2_E_6_3 (0)
    - Purchasing_Managers A_2_E_6_11 (0)

Cont'd

Professional_Occupations
Science_and_Technology_Professionals
Information_and_Communication_Technology_Professionals
Software_Professionals A_3_E_5_37, A_3_E_5_21 (0)
Science_Professionals
Physicists,_Geologists_and_Meterologists A_8_E_6_88 (0)
Engineering_Professionals
Electronics_Engineers A_8_E_6_24 (0)
Teaching_and_Research_Professionals
Research_Professionals A_4_E_14_22 (0)
Scientific_Researchers A_4_E_14_46 (0)
Teaching_Professionals
Registrars_and_Senior_Administrators_of_Educational_Establishments A_4_E_14_14 (0)
Primary_and_Nursery_Education_Teaching_Professionals A_4_E_14_78 (0)
Administrative_Occupations
Administrative_Occupations:_Records
Pensions_and_Insurance_Clerks A_8_E_5_88 (0)
Stock_Control_Clerks A_8_E_5_24 (0)

First 20 answers at dist.0 & 1 under all KAs

Associate_Professional_and_Technical_Occupations
- Science_and_Technology_Associate_Professional
  - IT_Service_Delivery_Occupations A_7_E_7_84 (0); **A_7_E_7_28 (1)**
    - IT_User_Support_Technicians A_1_E_8_98, A_1_E_8_58 (0); **A_7_E_7_12, A_1_E_8_50, A_7_E_7_68, A_1_E_8_74, A_1_E_8_34 (1)**
- Science_and_Engineering_Technicians **A_7_E_7_44 (1)**
  - Engineering_Technicians A_7_E_7_92 (0)
  - Quality_Assurance_Technicians A_7_E_7_4, A_7_E_7_60 (0); **A_7_E_7_60 (1)**
  - Laboratory_Technicians A_7_E_7_20 (0)
- Draughtspersons_and_Building_Inspectors
  - Architectural_Technologists_and_Town_Planning_Technicians A_7_E_7_52 (0)

Managers_and_Senior_Officials
- Corporate_Managers
  - Functional_Managers
    - Advertising_and_Public_Relations_Managers A_2_E_6_3 (0)
    - Purchasing_Managers A_2_E_6_11 (0)
    - **Personnel,_Training_and_Industrial_Relations_Managers A_2_E_6_51 (1)**
    - **Research_and_Development_Managers A_2_E_6_35 (1)**

Cont'd

Professional_Occupations
- Science_and_Technology_Professionals
  - Information_and_Communication_Technology_Professionals
    - Software_Professionals A_3_E_5_37, A_3_E_5_21 (0); **A_3_E_5_69 (1) A_8_E_5_96 (1)**
  - Science_Professionals
    - Physicists,_Geologists_and_Meterologists A_8_E_6_88 (0)
  - Engineering_Professionals
    - Electronics_Engineers A_8_E_6_24 (0)
    - **Civil_Engineers A_8_E_6_72 (1)**
    - **Mechanical_Engineers A_8_E_6_8 (1)**
    - **Planning_and_Quality_Control_Engineers A_8_E_6_40 (1)**

Teaching_and_Research_Professionals
- Research_Professionals A_4_E_14_22 (0)
  - Scientific_Researchers A_4_E_14_46 (0)
  - **Social_Science_Researchers A_4_E_14_6 (1)**
- Teaching_Professionals
  - Registrars_and_Senior_Administrators_of_Educational_Establishments A_4_E_14_14 (0)
  - Primary_and_Nursery_Education_Teaching_Professionals A_4_E_14_78 (0)

Cont'd

Administrative_Occupations

Administrative_Occupations:_Records

Pensions_and_Insurance_Clerks A_8_E_5_88 (0)

Stock_Control_Clerks A_8_E_5_24 (0)

Market_Research_Interviewers A_8_E_5_56 (1)

Library_Assistants/Clerks A_8_E_5_64 (1)

Transport_and_Distribution_Clerks A_8_E_5_16 (1)

Filing_and_Other_Records_Assistants/Clerks A_8_E_5_96 (1)

Extending Flexible Querying with KAs

We see that the relationships between the occupations returned as query results are now made explicit

In parallel, the user can explore increasingly larger fragments of the Occupation hierarchy, each rooted at a Knowledge Anchor that may be more meaningful to the user than a specialist occupation

This facilitates increasing awareness of possible relevant occupations by the user as compared with the purely linear presentation of results

Another Example Scenario

- Suppose the user wishes to become a Software Professional, or similar, and wishes to find out what subjects people have studied at university that have enabled them to get such a job
- This can be undertaken by evaluating the following SPARQL^{AR} query
- Before running the query, the user elects (through the system's GUI) to apply two of the edit operations available as part of the APPROX operator:
  - Insertion of an edge label; Substitution of an edge label;
- and also selects one relaxation operation from those available as part of the RELAX operator:
  - Replacement of a subclass by its immediate superclass;
- All edit/relaxation operations are set to have a cost of 1

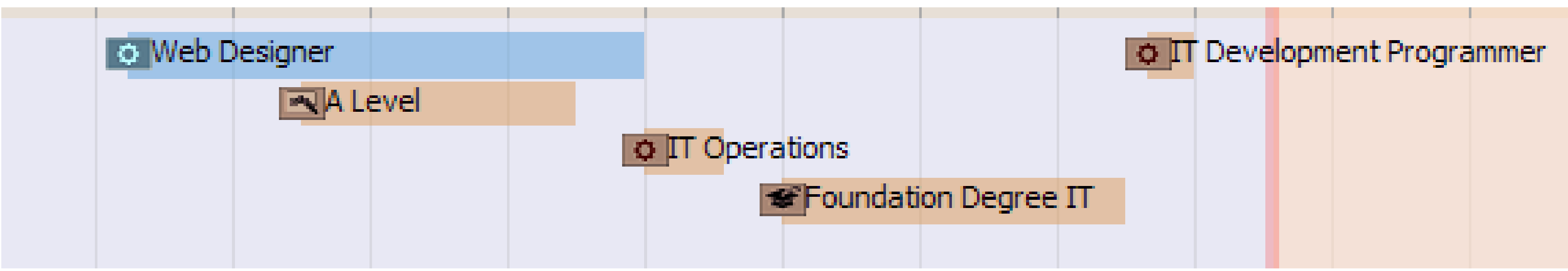
Extending Flexible Querying with KAs

```
SELECT ?UniEp ?Subj
WHERE {?UniEp  rdf:type  <http://www.L4All.com/University_Episode>.
      ?UniEp <http://www.L4All.com/qualif>/rdf:type ?Subj.
      APPROX( ?UniEp <http://www.L4All.com/prereq> ?WorkEp ).
      ?WorkEp  rdf:type  <http://www.L4All.com/Work_Episode>.
       RELAX ( ?WorkEp  <http://www.L4All.com/job>/rdf:type
                  <http://www.L4All.com/Software_Professionals> ) }
```

First 20 exact answers

A_3_E_4_15	Artificial_Intelligence
A_3_E_4_47	Artificial_Intelligence
A_8_E_3_45	Other_in_Mathematics_and_Computer_Sciences
A_8_E_3_62	Artificial_Intelligence
A_3_E_4_3	Statistics
A_3_E_4	Information_Systems
A_3_E_4_55	Artificial_Intelligence
A_3_E_4_97	Information_Systems
A_8_E_3_11	Operational_Research
A_3_E_4_71	Artificial_Intelligence
A_3_E_4_73	Information_Systems
A_3_E_4_7	Artificial_Intelligence
A_3_E_4_63	Artificial_Intelligence
A_3_E_4_25	Information_Systems
A_3_E_4_23	Artificial_Intelligence
A_3_E_4_69	Mathematics
A_3_E_4_35	Statistics
A_3_E_4_5	Mathematics
A_3_E_4_41	Information_Systems
A_3_E_4_49	Information_Systems

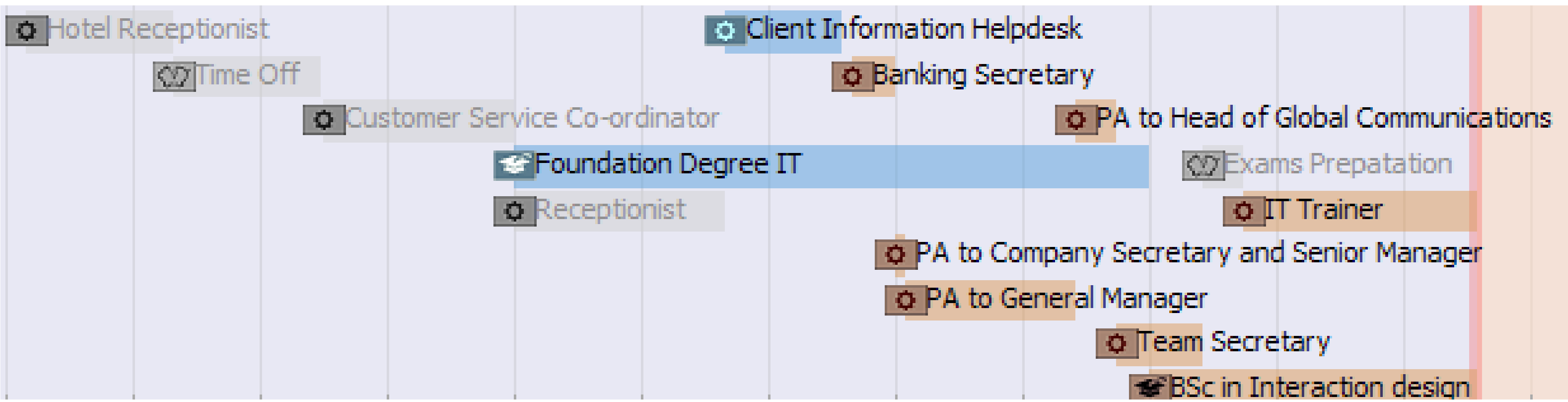
Clicking on an episode URI (in the first column of the results set) the user can explore a whole timeline



First 20 answers at distance 1

A_8_E_3_15	Software_Engineering
A_8_E_3	Information_Systems
A_8_E_3_83	Operational_Research
A_8_E_3_66	Statistics
A_8_E_3_49	Computer_Science
A_8_E_3_32	Information_Systems
A_1_E_5_75	Operational_Research
A_1_E_5_77	Other_in_Mathematics_and_Computer_Sciences
A_1_E_5_41	Computer_Science
A_1_E_5_53	Other_in_Mathematics_and_Computer_Sciences
A_1_E_5_83	Operational_Research
A_1_E_5_37	Other_in_Mathematics_and_Computer_Sciences
A_1_E_5_73	Computer_Science
A_1_E_5_25	Computer_Science
A_1_E_5_79	Software_Engineering
A_1_E_5_23	Software_Engineering
A_1_E_5_87	Software_Engineering
A_1_E_5_89	Computer_Science
A_1_E_5_51	Operational_Research
A_1_E_5_67	Operational_Research

Clicking on an episode URI (in the first column of the results set) the user can explore a whole timeline



First 20 answers at distance 2

A_1_E_3_75	Other_in_Mathematics_and_Computer_Sciences
A_1_E_5_48	Information_Systems
A_1_E_5_30	Artificial_Intelligence
A_1_E_3_77	Software_Engineering
A_1_E_5_24	Information_Systems
A_1_E_3_41	Operational_Research
A_1_E_5_40	Information_Systems
A_1_E_3_53	Software_Engineering
A_1_E_5_52	Mathematics
A_1_E_5_38	Artificial_Intelligence
A_1_E_5_42	Statistics
A_1_E_5_70	Artificial_Intelligence
A_1_E_3_83	Other_in_Mathematics_and_Computer_Sciences
A_1_E_5_94	Artificial_Intelligence
A_1_E_5_18	Statistics
A_1_E_3_37	Software_Engineering
A_1_E_5_82	Statistics
A_1_E_5_54	Artificial_Intelligence
A_1_E_5_74	Statistics
A_1_E_3_73	Operational_Research

First 20 answers at dist. 0, under KAs

Mathematical_and_Computer_Sciences

Artificial_Intelligence A_3_E_4_15, A_3_E_4_47, A_8_E_3_62, A_3_E_4_55, A_3_E_4_71, A_3_E_4_7, A_3_E_4_63, A_3_E_4_23 (0)

Information_Systems A_3_E_4, A_3_E_4_97, A_3_E_4_73, A_3_E_4_25, A_3_E_4_23, A_3_E_4_41, A_3_E_4_49 (0)

Mathematics A_3_E_4_69, A_3_E_4_5 (0)

Operational_Research A_8_E_3_11 (0)

Other_in_Mathematics_and_Computer_Sciences A_8_E_3_45 (0)

Statistics A_3_E_4_3, A_3_E_4_35 (0)

First 20 answers at dist. 0 & 1, under KAs

Mathematical_and_Computer_Sciences

Artificial_Intelligence A_3_E_4_15, A_3_E_4_47, A_8_E_3_62, A_3_E_4_55, A_3_E_4_71, A_3_E_4_7, A_3_E_4_63, A_3_E_4_23 (0)

Computer Science A_8_E_3_49, A_1_E_5_41, A_1_E_5_73, A_1_E_5_25, A_1_E_5_89 (1)

Information_Systems A_3_E_4, A_3_E_4_97, A_3_E_4_73, A_3_E_4_25, A_3_E_4_23, A_3_E_4_41, A_3_E_4_49 (0), **A_8_E_3, A_8_E_3_32 (1)**

Mathematics A_3_E_4_69, A_3_E_4_5 (0)

Operational_Research A_8_E_3_11 (0), **A_8_E_3_83, A_1_E_5_75, A_1_E_5_83, A_1_E_5_51, A_1_E_5_67 (1)**

Other_in_Mathematics_and_Computer_Sciences A_8_E_3_45 (0), **A_1_E_5_77, A_1_E_5_53, A_1_E_5_37 (1)**

Software Engineering A_8_E_3_15, A_1_E_5_79,A_1_E_5_23,A_1_E_5_87 (1)

Statistics A_3_E_4_3, A_3_E_4_35 (0), **A_8_E_3_66 (1)**

First 20 answers at dist. 1 & 2, under KAs

Mathematical_and_Computer_Sciences

Artificial_Intelligence **A_1_E_5_30, A_1_E_5_38, A_1_E_5_70, A_1_E_5_94, A_1_E_5_54 (2)**

Computer Science A_8_E_3_49, A_1_E_5_41, A_1_E_5_73, A_1_E_5_25, A_1_E_5_89 (1)

Information_Systems A_8_E_3, A_8_E_3_32 (1), **A_1_E_5_48, A_1_E_5_24, A_1_E_5_40 (2)**

Mathematics **A_1_E_5_52 (2)**

Operational_Research A_8_E_3_83, A_1_E_5_75, A_1_E_5_83, A_1_E_5_51, A_1_E_5_67 (1), **A_1_E_3_41, A_1_E_3_73 (2)**

Other_in_Mathematics_and_Computer_Sciences A_1_E_5_77, A_1_E_5_53, A_1_E_5_37 (1), **A_1_E_3_75, A_1_E_3_83 (2)**

Software Engineering A_8_E_3_15, A_1_E_5_79,A_1_E_5_23,A_1_E_5_87 (1), **A_1_E_3_77, A_1_E_3_53, A_1_E_3_37 (2)**

Statistics A_8_E_3_66 (1), **A_1_E_5_42, A_1_E_5_18, A_1_E_5_82, A_1_E_5_74 (2)**

Conclusions and Future work

This work addresses the challenge of supporting the exploration of large knowledge graphs by users who are not experts in the domain

We have proposed an approach combining flexible graph querying and knowledge anchors:

- flexible queries allow automatic expansion of query results by query approximation and query relaxation
- KAs represent basic-level entities close to the user's cognitive structures; they are likely to be familiar to many users and can provide good starting points for introducing unfamiliar entities

Conclusions and Future work

In our hybrid approach, we introduce knowledge anchors into query results by including *paths to the nearest knowledge anchor(s)*

This facilitates increasing awareness of possible relevant Occupations and Subjects by the user as compared with the purely linear presentation of results

Not only can this facilitate exposing learning and work possibilities that may not otherwise have been considered, but it can also help users to more accurately classify episodes within their *own* timeline, thereby increasing the quality of the search results for others

Conclusions and Future work

Validation studies have been recently undertaken comparing the KA metrics against users' own perceptions of basic-level objects in the L4All ontology

see Al-Tawil et al, Evaluating Knowledge Anchors in Data Graphs Against Basic Level Objects, ICWE 2017, pp 3-22

with generally promising results and several recommendations on how to best combine the 6 metrics

Future work involves development of visualisations such as those illustrated here and evaluation with groups users to compare the effectiveness of the alternative forms of query results presentation

It would also be interesting to investigate other ways of hybridising flexible queries and knowledge anchors, e.g. for filtering or ranking query results

QUESTIONS?

User Study on Basic Level Objects

For details, see M. Al-Tawil et al,

Evaluating Knowledge Anchors in Data Graphs Against Basic Level Objects, Proc. ICWE 2017, pp 3-22

Participants:

- 28 participants, university students and professionals, age 25–64, recruited on a voluntary basis. Most of them were experienced mainly in Computing.

Category identification task:

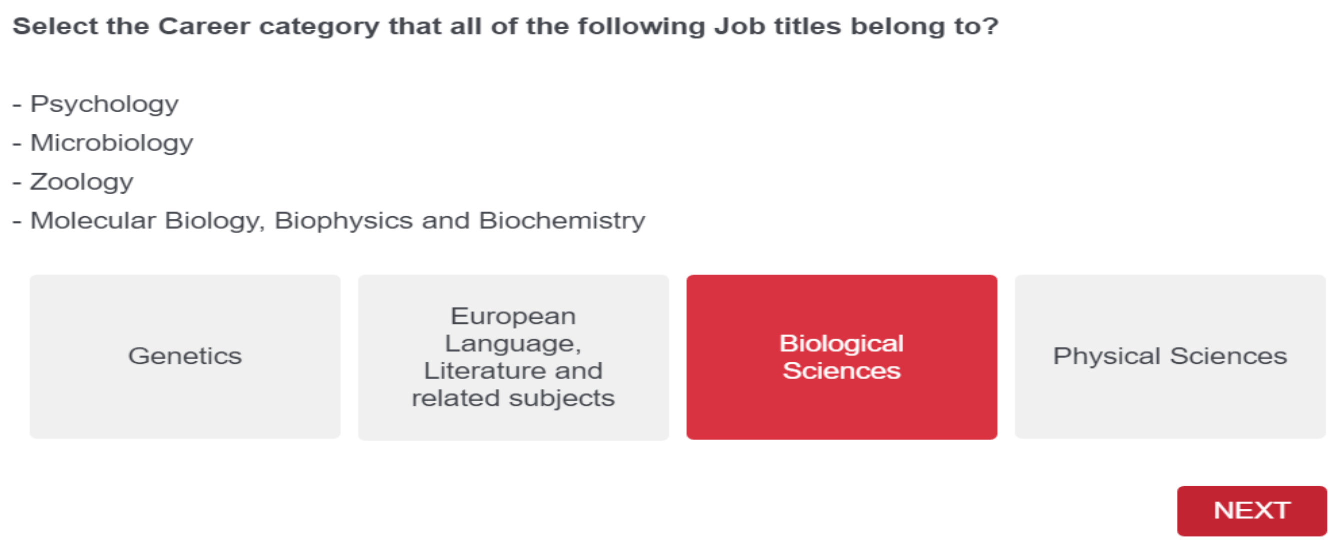
- Seven online surveys were developed (6 presented the 114 category entities of the *Occupation* class hierarchy, with each survey showing 19 categories; one survey presented the 19 categories of the *Subject* class hierarchy). The category allocation in each survey was random. Every survey had 4 respondents from the study participants. Each participant was allocated only to one survey.
- A representation of each category was shown on the participant's screen and he/she was asked to identify the category name.
- The representation included a list of leaves’ names of that category (at most four leaf names were shown on the participant's screen).

- The participant was provided with four different categories as candidate answers and the participant was asked to select one category that he/she thinks the leaf entities belong to:

    - the category all the leaves belong to
    - a parent from the superordinate level
    - a member from the subordinate level
    - a sibling at the same category level

- In cases where no parents or members could be added to the candidate answers, siblings were used instead.
- Figures 5 and 6 show examples of the category identification task from the *Occupation* and *Subject* class hierarchies respectively.

- The participant in Figure 5 saw two leaves (the category has two leaves only) of the category Housekeeping Occupation and the participant identified the category's parent Personal Service Occupation, which he/she thinks that the leaves belong to. This will increase the count for the category Personal Service Occupation.



- In Figure 6, a participant was shown leaf names of the category Biological Sciences (four random leaves where selected among 9) and selected its exact name. This will increase the count for the category Biological Sciences.

Category entities in the *Occupation* and *Subject* class hierarchies that were named by at least two different users were identified as Basic Level Objects (BLOs)

The full Knowledge Anchors (KA) and BLO sets obtained from the L4All data set are available here:

https://drive.google.com/drive/folders/0B5ShywKndSLXaVhrSWpiYVZ3WjA

Hybridization:

- Analysis of the False Positive and False Negative entities in the KA sets indicated that the metrics had different performance on different taxonomical levels in the L4All data graph, which is captured in these two heuristics:

- *Heuristic 1: Use the AV and CAC distinctiveness metrics with hierarchical relationships for the categories at the bottom quartile of the class taxonomy.*
- *Heuristic 2: Use majority voting for all other taxonomical levels.*

- This hybridization increased the performance of the KA metrics to:
  - for *Occupation*, Precision = 0.77 and Recall = 0.92
  - for *Subject*, Precision = 1 and Recall = 0.53