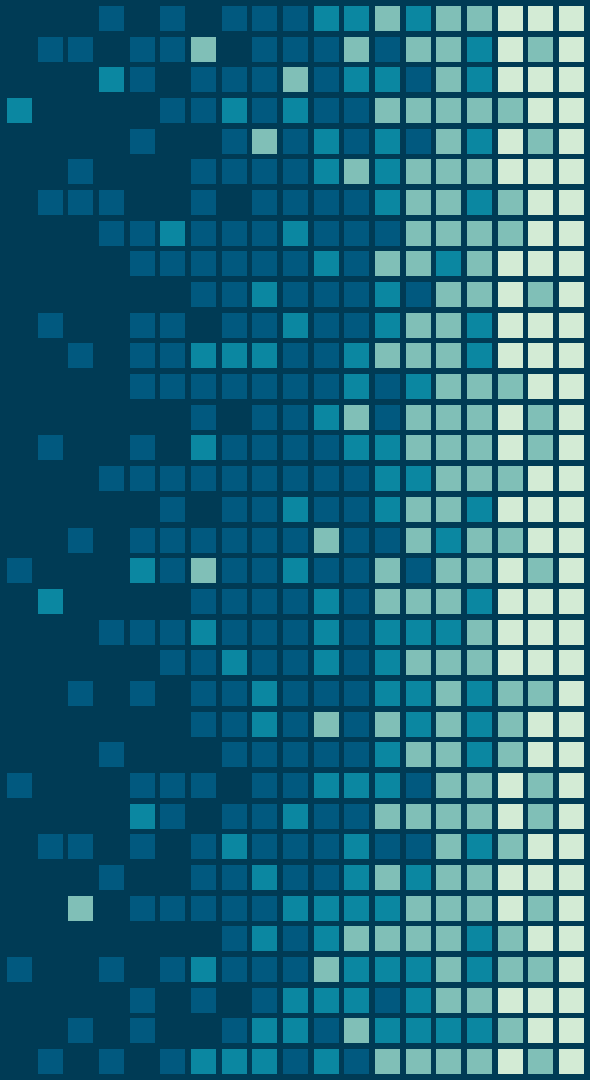
AquaCold

Aggregated **Q**uery **U**nderstanding **A**nd **C**onstruction **O**ver **L**inked **D**ata

Nick Collis & Ingo Frommholz
University of Bedfordshire



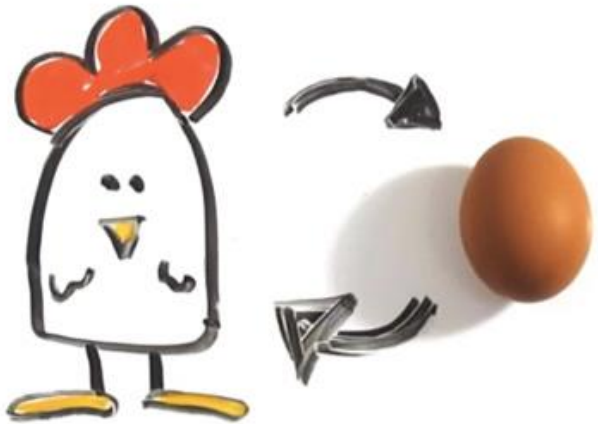
The Linked Data Web today

The Linked Data Web has existed for over 15 years and grown to incorporate **over 149 billion triples**.

Previous problem - Not enough useful linked open data.

Current problem - Now lots of useful linked data but how can we search and explore it effectively?

Then: 'Chicken and egg' problem



Now: *"We have enough chickens, but how can we get enough eggs out of them?"**

*** S.Ferre**, Expressive and Scalable Query-based Faceted Search over SPARQL Endpoints, 2014

Difficulties with searching linked data

SPARQL remains the most popular method for querying Linked Data. But…

```
PREFIX dbo:
<http://dbpedia.org/ontology/>
PREFIX dbp:
<http://dbpedia.org/resource/>
PREFIX foaf:
<http://xmlns.com/foaf/0.1/>

SELECT ?name ?bandname where
{
?person foaf:name ?name .
?band dbo:bandMember ?person .
?band dbo:genre dbpedia:Punk_rock .
?band dbp:name ?bandname .
}
```

SPARQL syntax is unintuitive for non technical users.

The SPARQL endpoint must be known in order to write queries.

Users must understand the **RDF model**, and **ontology terms**

Many instances of ambiguity in data (eg :Place vs PopulatedPlace)

"...the lack of technical knowledge and an understanding of the intricacies of the semantic technology stack limits users in their ability to interpret and make use of the Web of Data....

...the key solution is to visualise Linked Data in a coherent and legible manner, allowing non-domain and non-technical audiences to obtain an understanding of its structure, and therefore implicitly compose queries, identify links between resources and intuitively discover new pieces of information"

Dadzie, A.s., Rowe, M.: Approaches to Visualising Linked Data: A Survey
www.semantic-web-journal.net.

Goals

Make linked data search more usable:

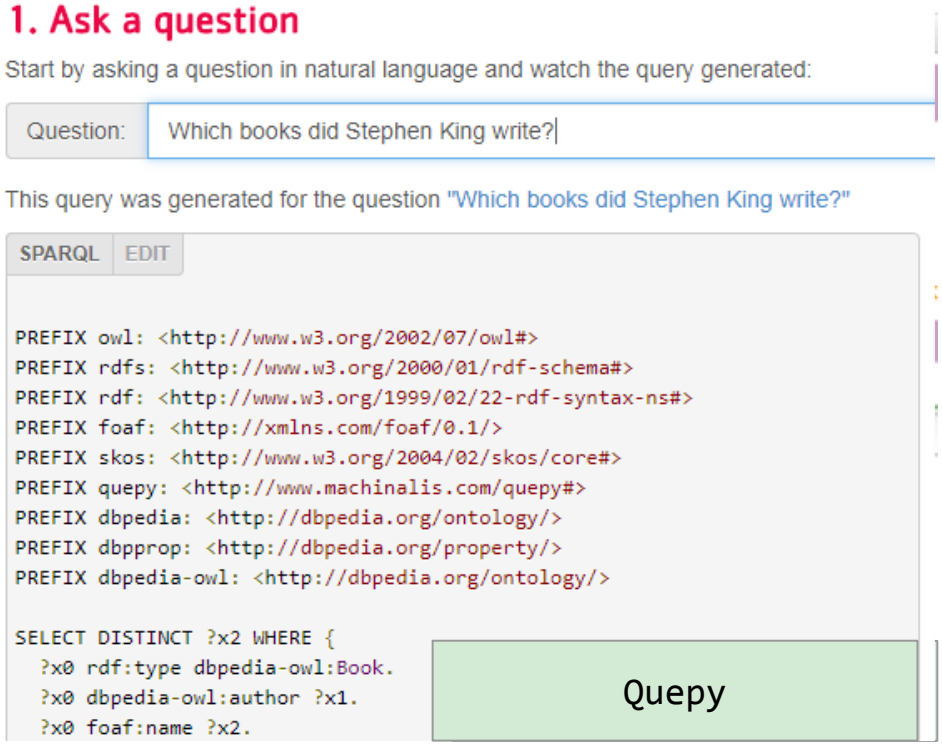
- **Abstract the complexity** of SPARQL whilst retaining expressivity
- Improve the **discoverability** of endpoints and ontologies
- Provide a measurement of **result accuracy**
- Provide a way of identifying **ambiguous** labels
- Build a **scalable** system, able to cope with vast amounts of data

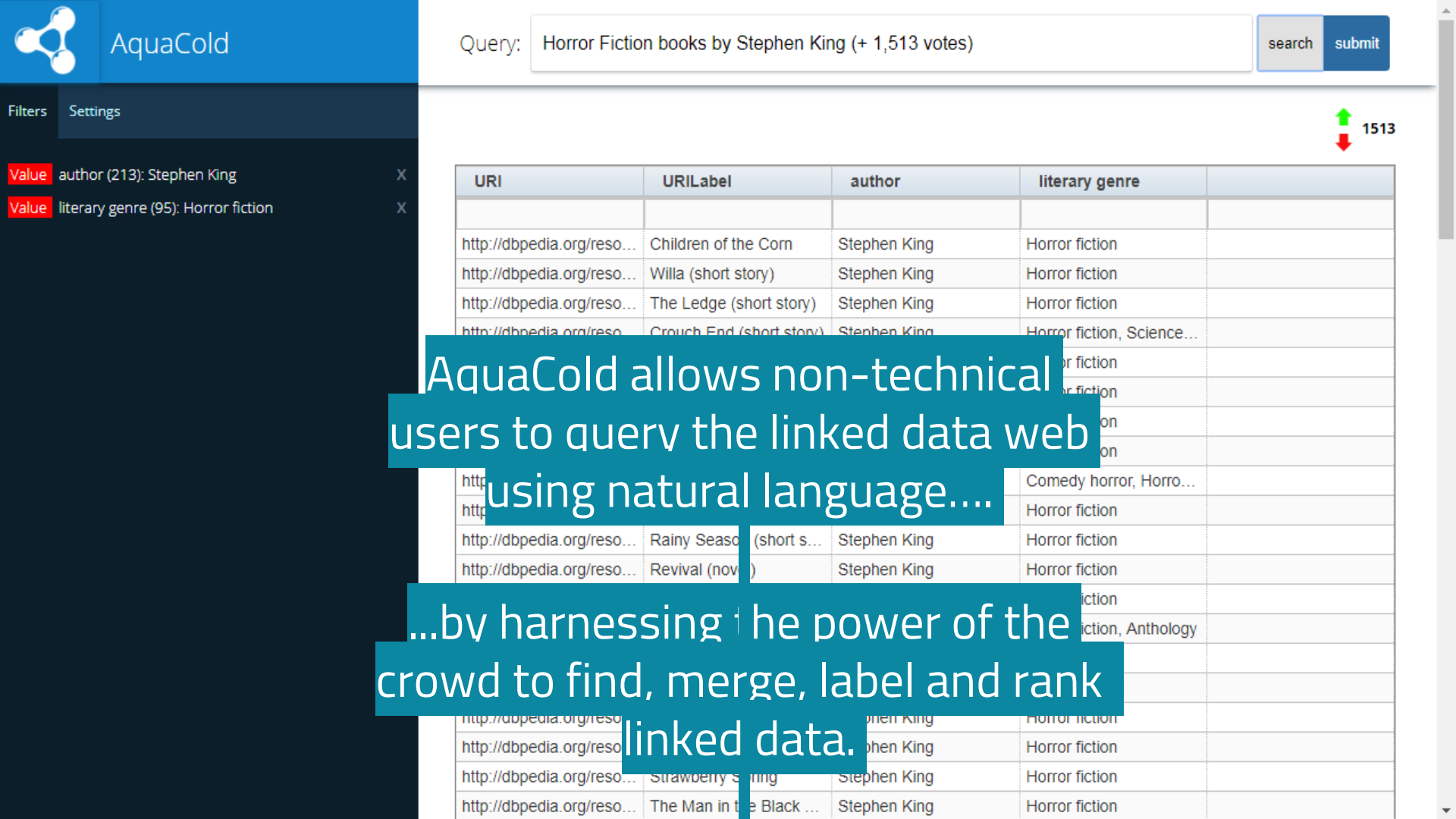
Approaches to improving LD search usability

Graphical Query Builders
+ More readable than SPARQL
- Less intuitive than NL
- Lacks expressivity

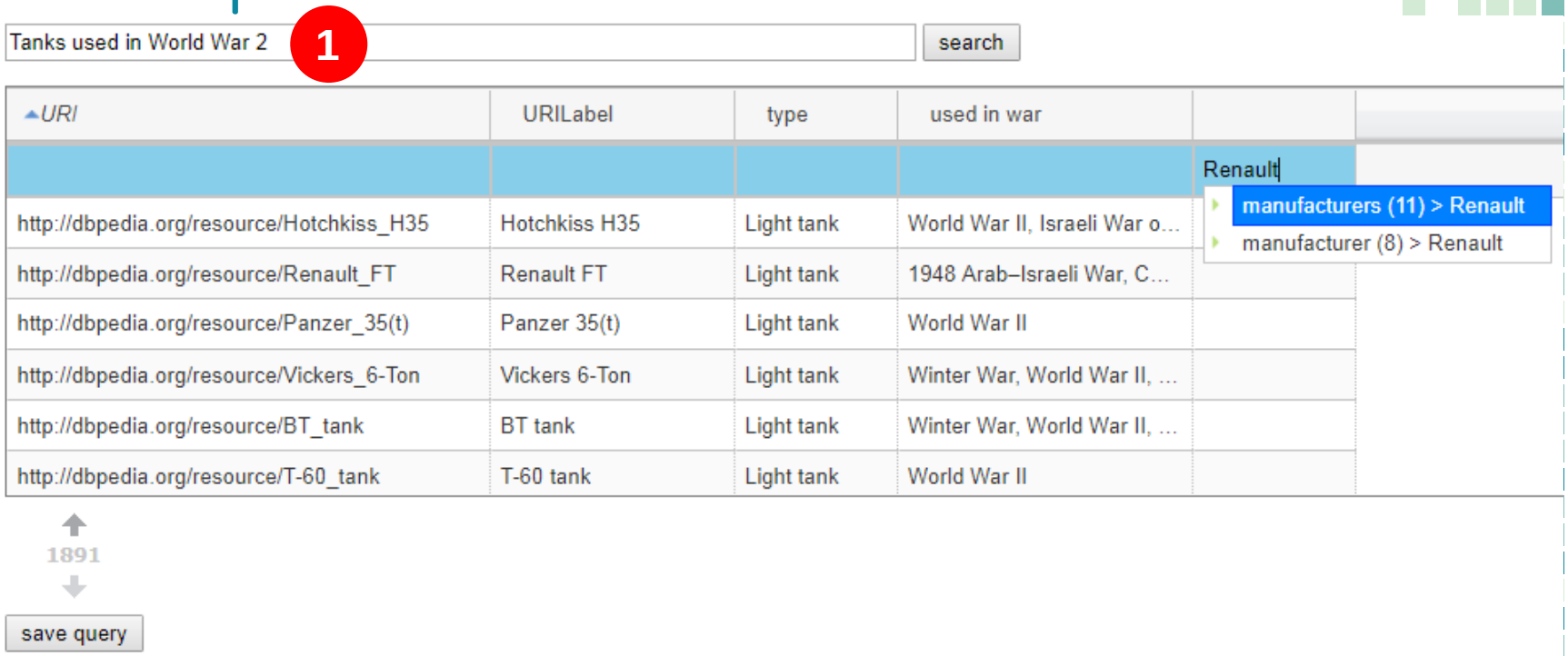
Controlled Natural Language
+ Readable, intuitive, robust
- Inflexible

Uncontrolled Natural Language
+ Highly intuitive when working
- Less accurate results, particularly for complex queries



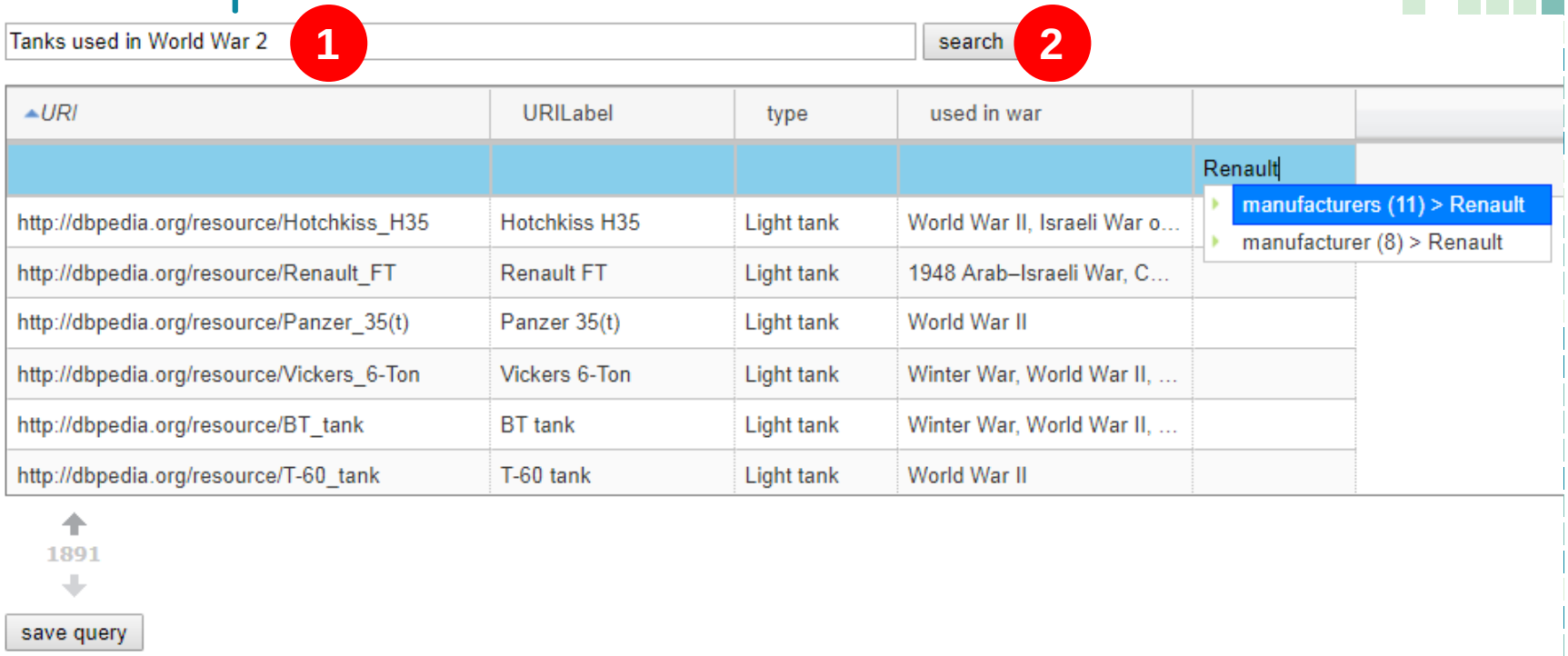

AquaCold
Query:
Horror Fiction books by Stephen King (+ 1,513 votes)
search
submit
Filters
Settings
Value author (213): Stephen King
Value literary genre (95): Horror fiction
1513
URI
URILabel
author
literary genre
http://dbpedia.org/reso…
Children of the Corn
Stephen King
Horror fiction
http://dbpedia.org/reso…
Willa (short story)
Stephen King
Horror fiction
http://dbpedia.org/reso…
The Ledge (short story)
Stephen King
Horror fiction
Horror fiction, Science…
Comedy horror, Horro…
Horror fiction
http://dbpedia.org/reso…
Rainy Seaso (short s…
Stephen King
Horror fiction
http://dbpedia.org/reso…
Revival (nov)
Stephen King
Horror fiction
iction, Anthology
http://dbpedia.org/reso…
Strawberry Spring
Stephen King
Horror fiction
http://dbpedia.org/reso…
The Man in t e Black …
Stephen King
Horror fiction
AquaCold allows non-technical users to query the linked data web using natural language....
...by harnessing the power of the crowd to find, merge, label and rank linked data.

The Aquacold interface



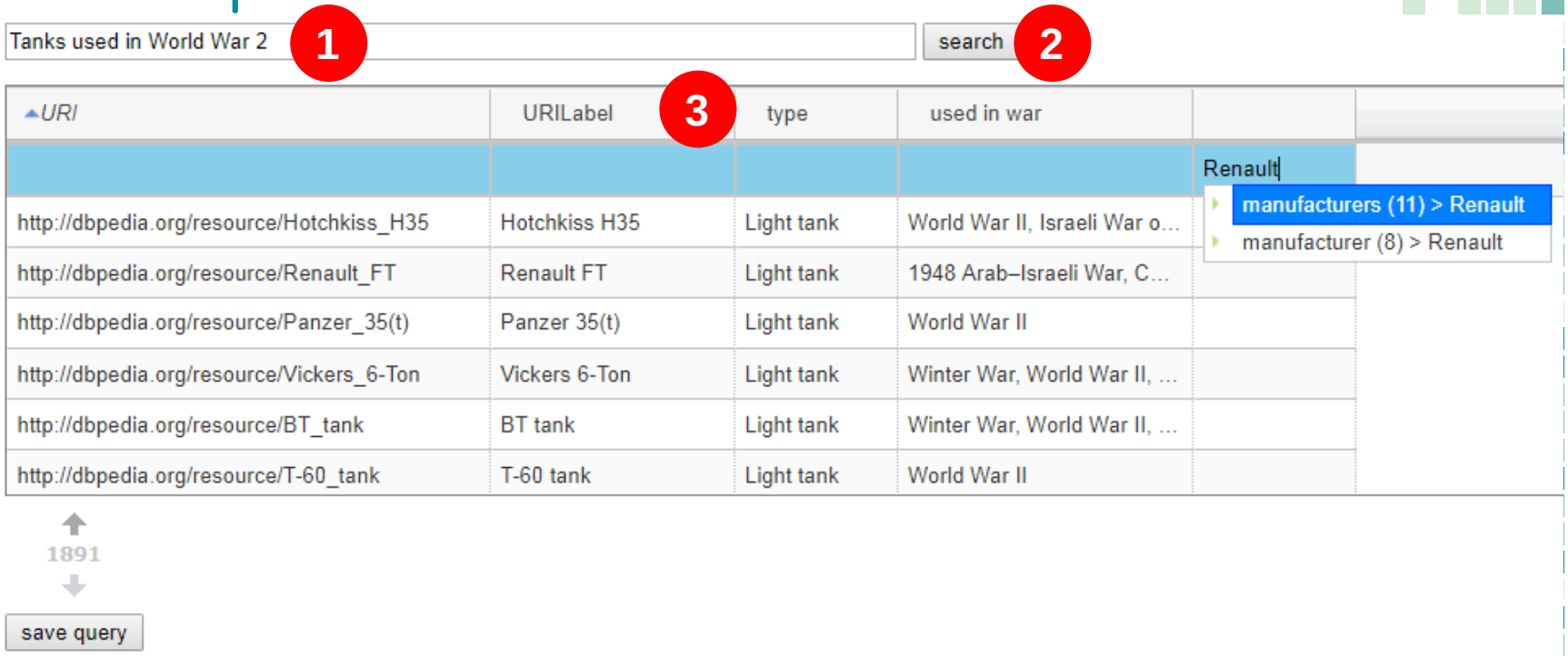
1 **Search & Labelling box** – Users type NL query here and also label the results grid using NL

The Aquacold interface



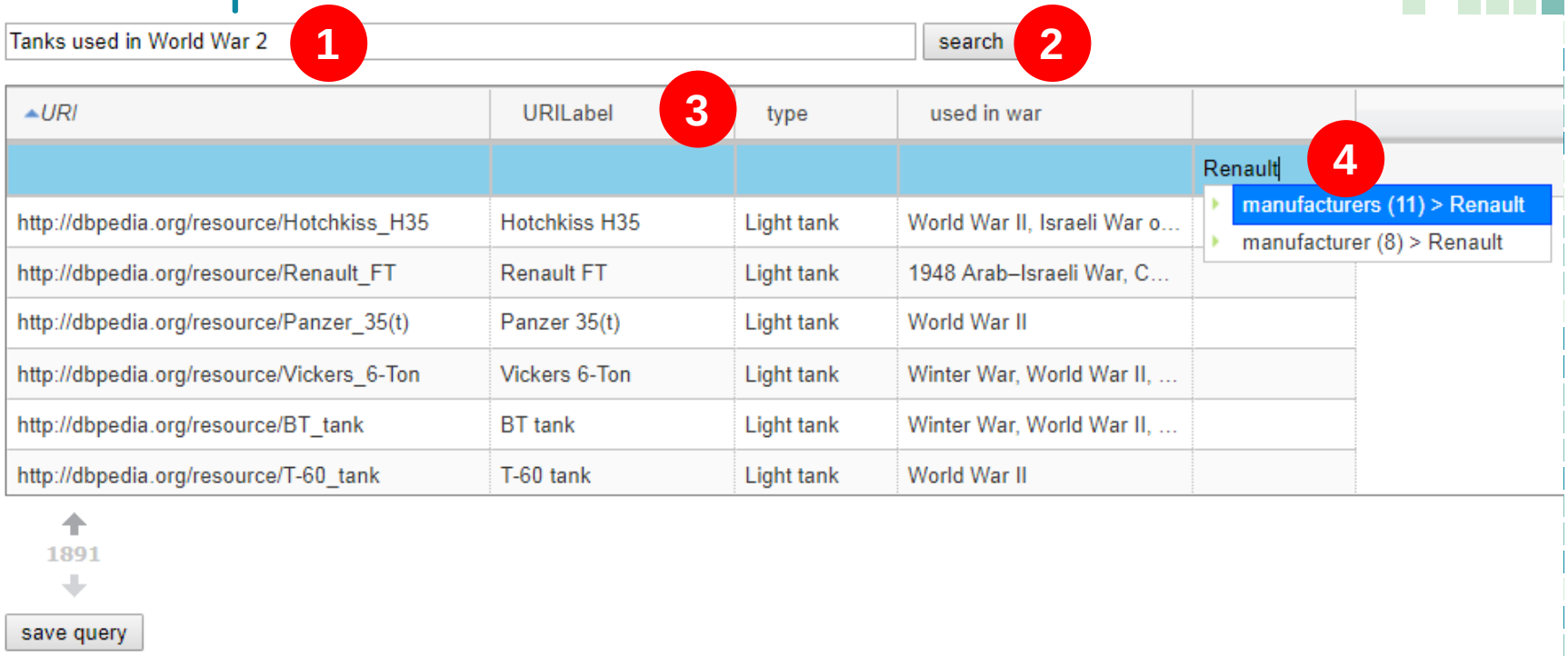
2 **Search button** – On clicking this, the results grid will be refreshed with the results of the query.

The Aquacold interface



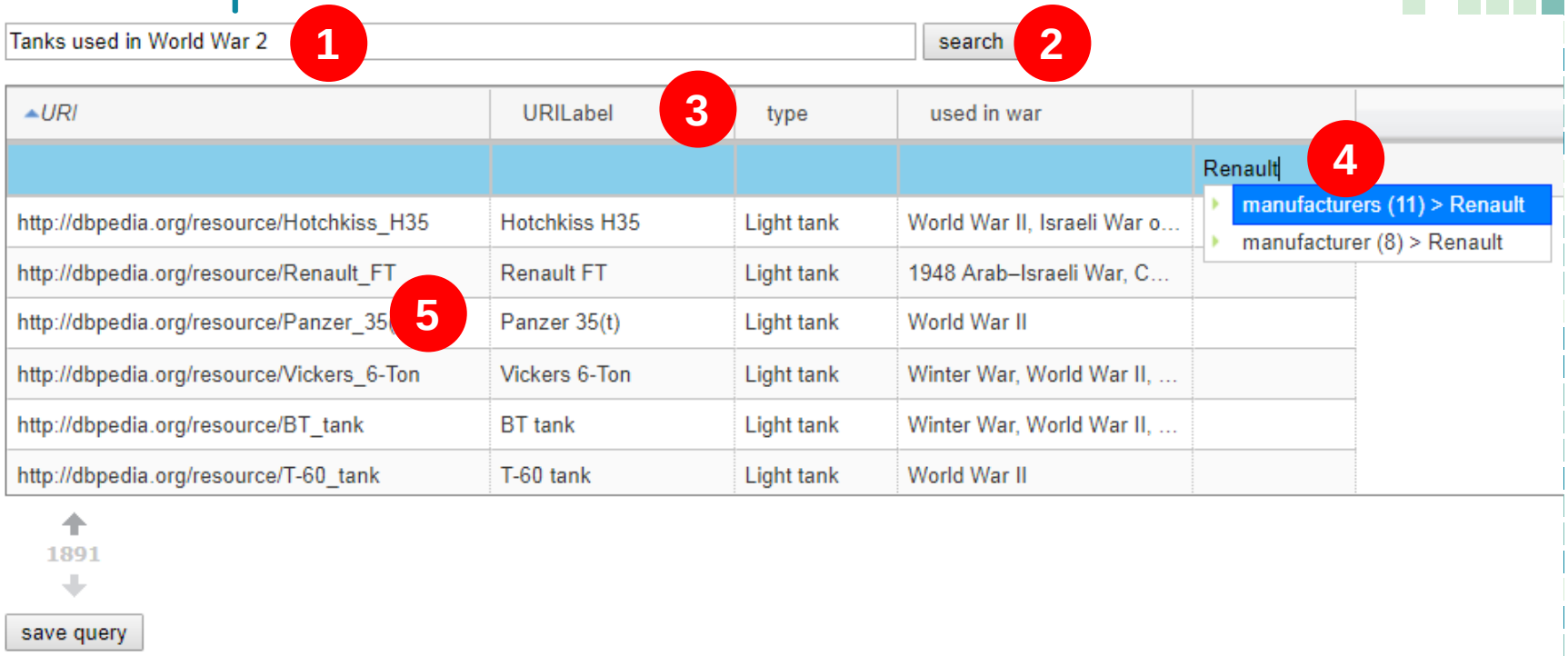
3 **Property filters** – Used to search for properties of an object, such as type, age, weight, etc..

The Aquacold interface



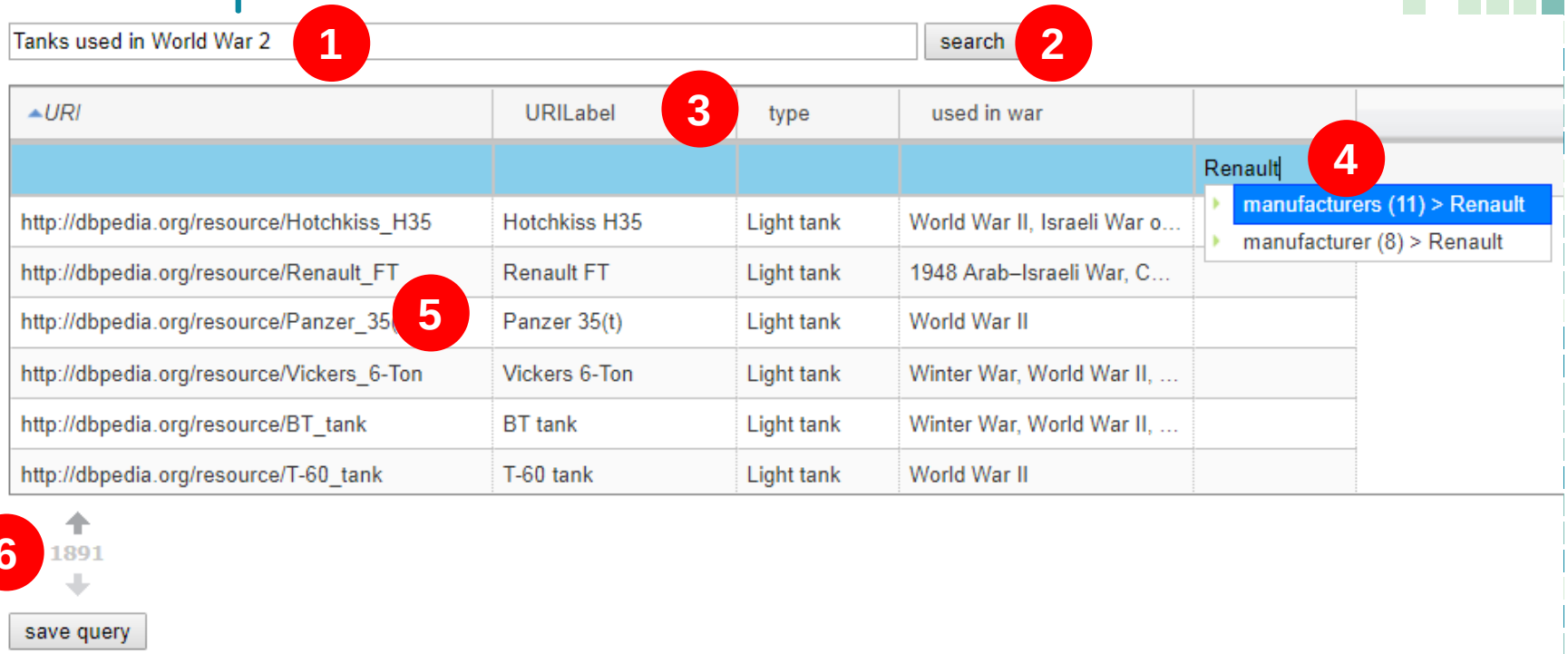
4 **Subject & property filters** – Used to search for a property / value pair.

The Aquacold interface



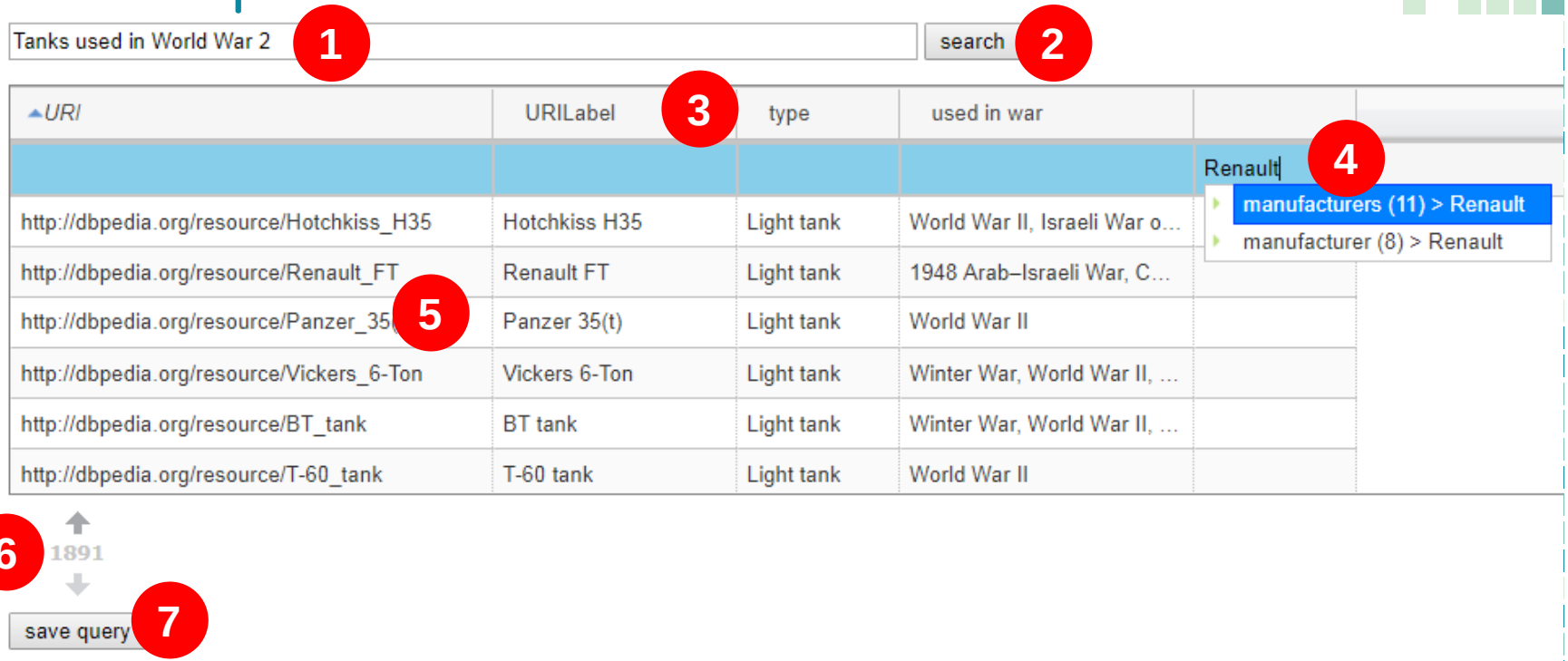
5 **Results grid –** Lists the results for this query. These can be manipulated using the filters.

The Aquacold interface



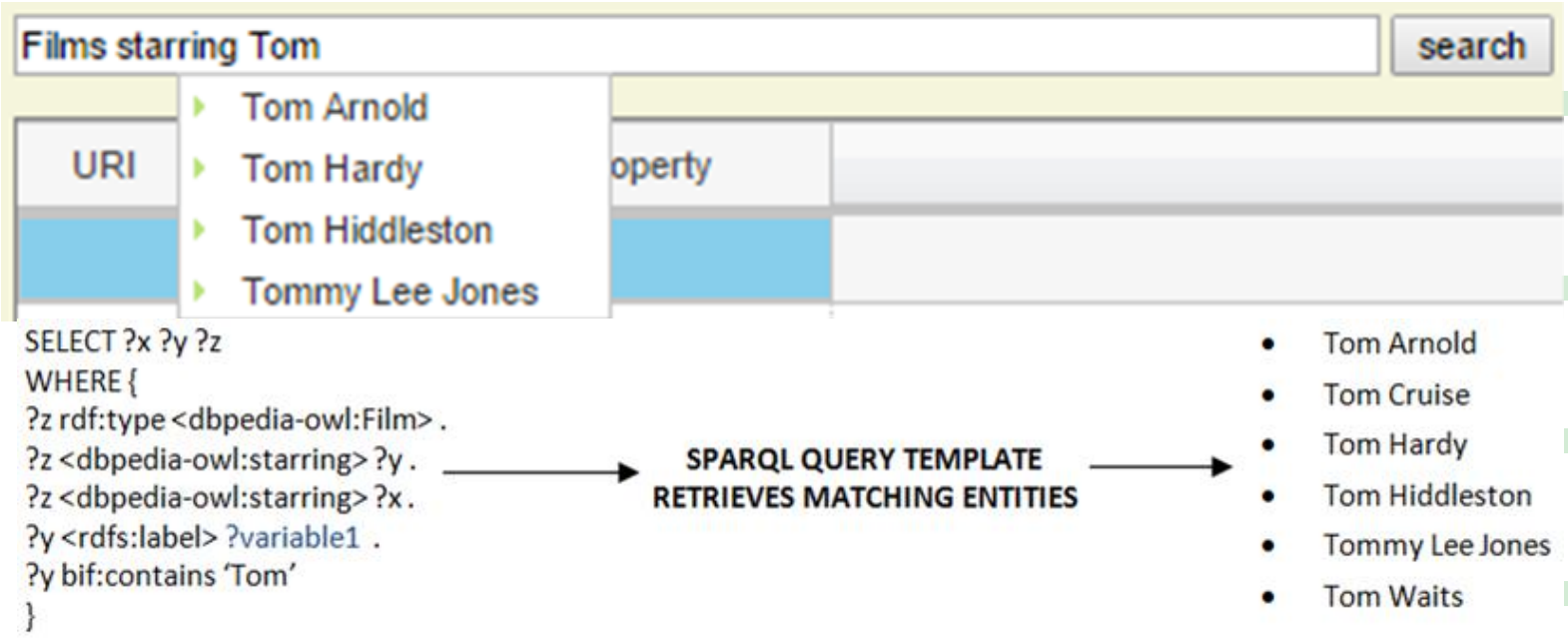
6 **Results grid voting** – Users vote on the accuracy of the results and label using these controls.

The Aquacold interface



7 **Save query button** – This saves the query to the database with the label defined in box 1.

AquaCold – Querying



Autocomplete suggestions are based on labels written by other users

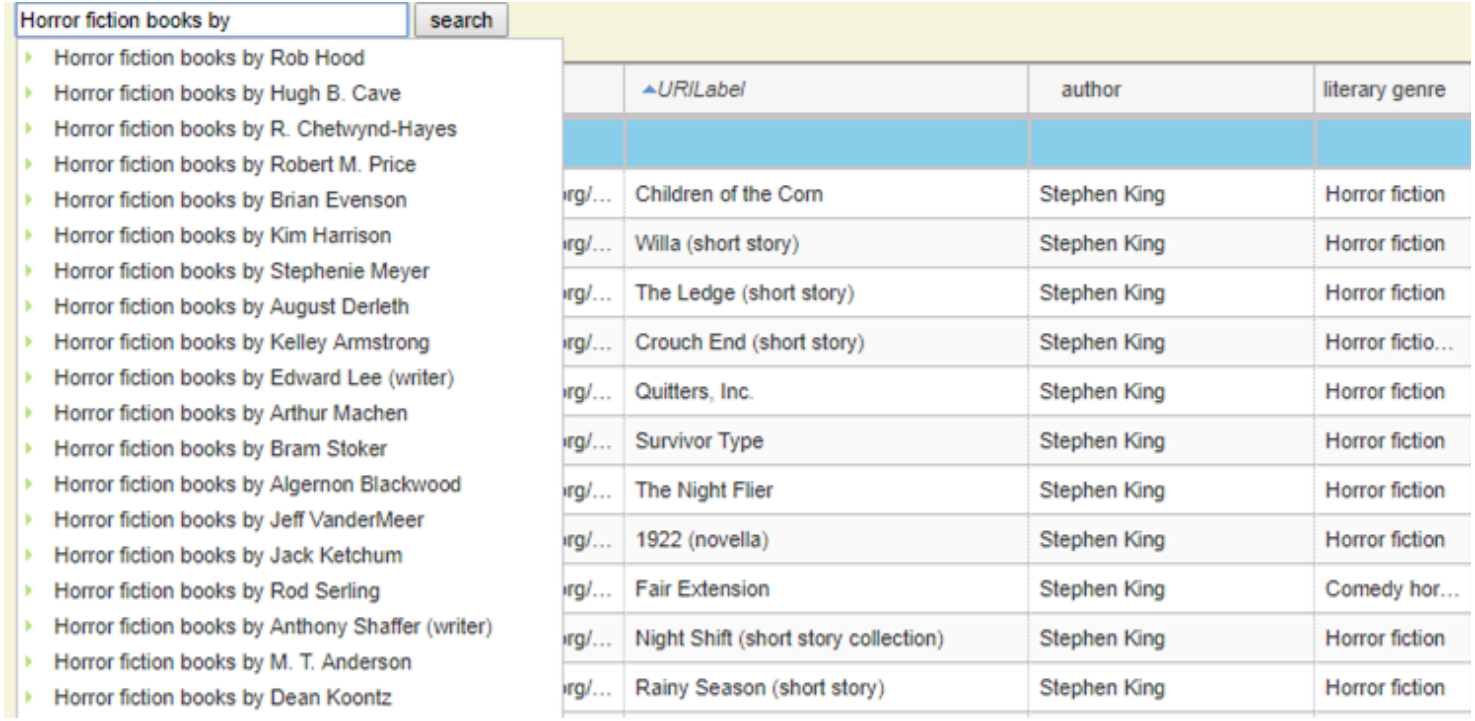
AquaCold – Building result grids

```
SELECT ?property ?y
WHERE {
?x ?property ?y .
?y rdf:label ?z .
FILTER regex(?z, "Al Pac", "i")
}
```

URI	URILabel	starring
http://dbpedi...	...And Justice for All (film)	John Forsythe, John Forsythe, Lee Strasberg, Lee Strasberg, Al Paci
http://dbpedi...	Author! Author! (film)	Tuesday Weld, Dyan Cannon, Bob Dishy, Al Pacino, Alan King (comec
http://dbpedi...	Glengarry Glen Ross (film)	Jack Lemmon, Kevin Spacey, Alan Arkin, Jonathan Pryce, Al Pacino, A

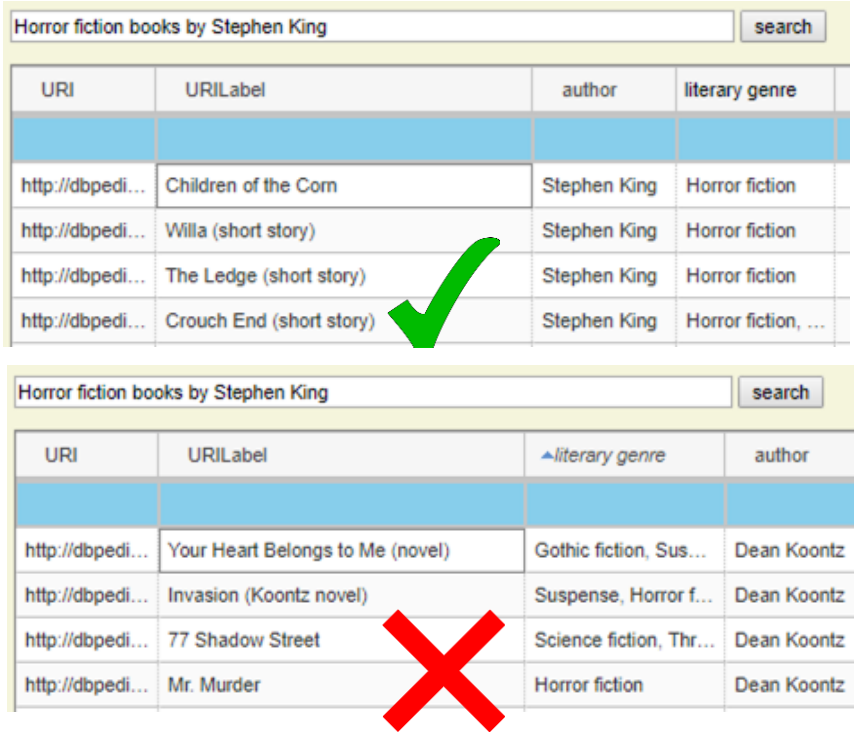
Entity autocompletion is carried out using basic regex string matching

AquaCold – Labelling Linked Data Sets



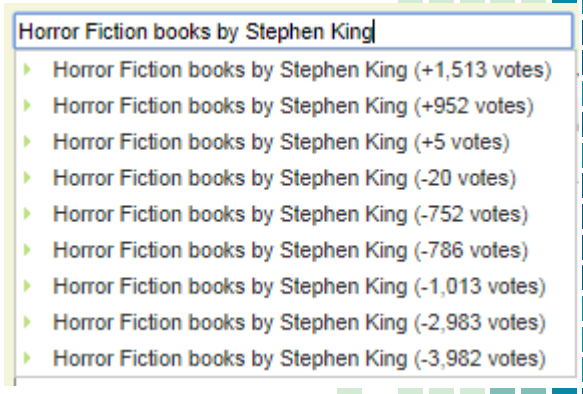
Result grids can be labelled with NL, guided by labels used by others.

AquaCold - Voting



User Score

+1,513

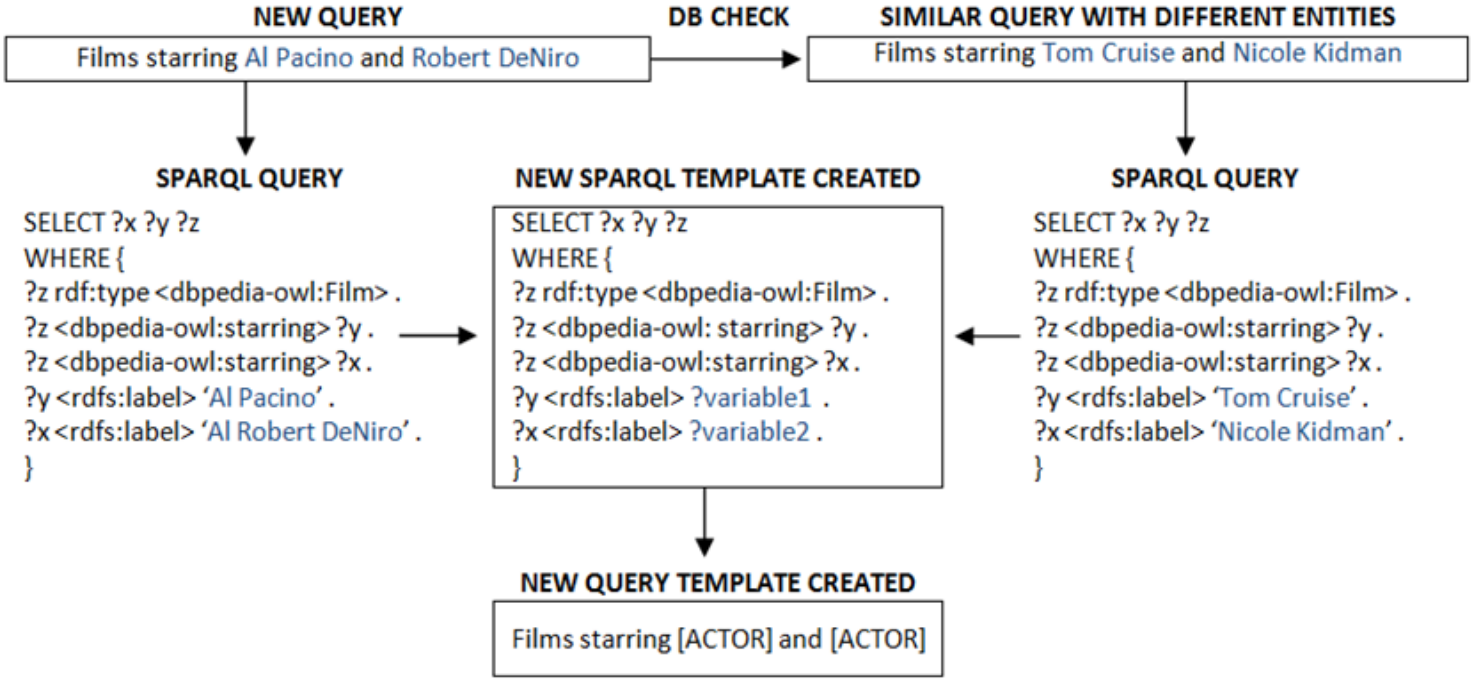


User Score

-2,983

Uses crowdsourced voting on labels to surface high quality results

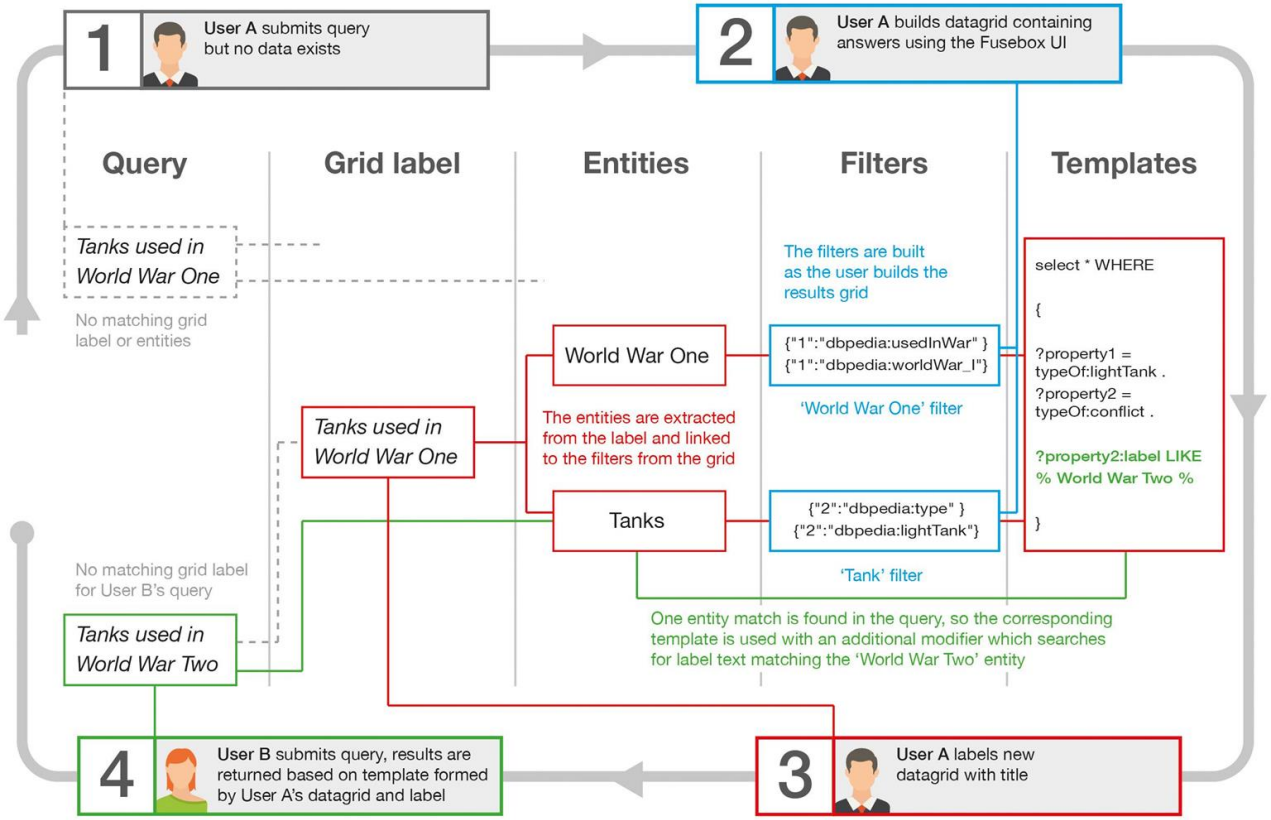
AquaCold – Templating



Creates templates based on labels to seed multiple label variants.

Demonstration

[Video 1](#) [Video 2](#) [Video 3](#) [Video 4](#)



User A

Looking for *Tanks used in World War One*

User B

Looking for *Tanks used in World War Two*

[site](#)

How AquaCold allows usable Linked Data Search

Abstract SPARQL complexity **retaining expressivity**

- Natural language input abstracts complexity of SPARQL for querying
- Faceted results grid retains some SPARQL expressivity for exploration

Improve discoverability of *endpoints* and *ontologies*

- Users discover endpoints, build result grids and label for other users

Providing a **measurement of accuracy** for results

- Crowdsourced voting tools surface the most accurate results

Enabling **disambiguation** of result labels

- Multiple identical labels are allowed (attached to different result grids)
- Voting mechanism surfacing the most relevant

More benefits of AquaCold approach

- Natural language queries for linked data result sets.
  - Highly intuitive for non technical users. Allows synonyms, slang etc.
- Guided labelling assistance.
  - Displays similar existing labels to limit label ambiguity.
- Faceted interface for composing result sets.
  - Hides SPARQL complexity. No need to know the endpoint details.
- Voting mechanism for labels and result sets.
  - No need to know the endpoint, ontology terms or URIs.
- Template label generation.
  - Ameliorates cold start problem

Limitations of AquaCold approach

- Basic voting system open to misuse
  - Simple +1, -1 voting could be easily manipulated.
- Grid based interface suits some queries more than others
  - Queries with many links between nodes. Eg FOAF less suited
- Could lead to issues with ambiguity in query labels.
  - Disparity and ambiguity in LD could transfer to AquaCold
- System is very early in development
  - Many features not yet present, eg: AND, NOT, node linking
- Lacks expressivity compared to SPARQL
  - Cannot match the expressivity of dedicated query language

Evaluation planned for early 2018

Evaluation will use questions from the QALD challenge

Established benchmark used by other NL systems inc. SPARKLIS, TBSL

Measured against similar NL LD search systems

Including SPARKLIS, CrowdQ, Quepy, Ginseng, Semantic Crystal

Key challenge - parity with other systems

Allow accurate comparisons with other systems taking into account features that are unique to AquaCold

Test groups to have a range of experience levels

Will measure **Expressivity**, **Scalability** and **Usability**

Future work

Hypothesis

human labelled result sets and crowdsourced quality control will result in higher precision compared to systems which use programmatic translation of natural language queries to retrieve linked data.

Enhance crowdsourcing & voting system:

Investigate literature for enabling effective crowdsourced voting

Build more advanced filters

NOT, AND, numeric < >, date operators, fuzzy regex matching….

Enhance entity detection

Investigate entity detection tools such as IBM Watson or Word2Vec

THANKS!

Any questions?

You can find me at:
nwcphd@gmail.com