A New Approach to Impact Case Study Analytics

[1] **Jiajie Zhang,** [2] **Paul Watson*,** [3] **Barry Hodgson**

[1,2,3]*School of Computing, Newcastle University,Newcastle upon Tyne, UK*
[2]*The Alan Turing Institute, London, UK*
paul.watson@newcastle.ac.uk

Abstract

The 2014 Research Excellence Framework (REF) assessed the quality of university research in the UK. A fifth of the assessment came from peer review of the impact of research outside academia, reflecting its growing importance in UK government policy. Impact is defined as an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life. Each university submitted a set of four-page impact case studies for assessment. These are mainly free text that describes and evidences the impact of research. There have been several analyses of these case studies, but these have either used qualitative methods or basic forms of text searching and analysis. These approaches have limitations, especially in terms of the time needed to analyse the data manually, and due to the often poor quality of the answers generated by applying computational analysis to free text data that lacks structure and context. This paper describes a new system we have built that that takes an alternative approach to overcome these problems. At its core is a structured, queryable representation of the Impact Case Study data. We describe the design of the ontology used to structure the information, and how semantic web technologies are used to store and query the data. We show that this gives two main advantages compared to existing techniques: improved accuracy in question answering, and the ability to answer a broader range of questions, including by integrating data from external sources.

Keywords— Infrastructure; Knowledge Generation; Semantic Web; Ontology; Impact.

1 Introduction

The 2014 Research Excellence Framework (REF) was a peer assessment of the quality of UK universities research in all disciplines. One-fifth of the overall assessment was allocated by peer review of the impact of research. The impact is defined as an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia[1]. Each university submitted a set of four-page impact case studies for each unit of assessment. These describe and evidence the impact of research conducted by one or more academics.

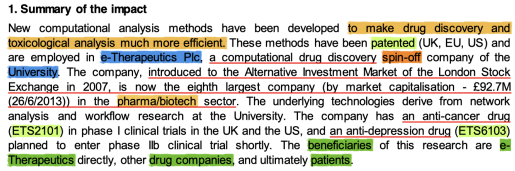
The open publication of all the case studies has enabled analysis to gain insights and answer questions about them. However, they are mainly free text, and the lack of structure means that this has to date been an exercise in qualitative analysis or text search. For example, one of the most comprehensive studies comes from King's College London and Digital Science Grant (2015); it includes word frequencies and the breakdown of impact topics. The main methods they used were topic modelling, named entity recognition and term frequency. However, there are specific limitations — in particular, while it is possible to find terms representing individuals, institutions and companies from the case studies, this does not in itself reveal the context in which they appear, and therefore the meaning and significance. For example, a text search may show that "IBM" frequently occurs in the case studies, but this does not tell us about its role, e.g. collaborator, research output user, customer, investor, funder, or spin-out acquirer.

The goal of our research was therefore to address these limitations by exploring an alternative approach. We designed and built a system that has at its heart a structured, queryable representation of the data found in Impact Case Studies. The case studies are transformed into a structured format according to a custom-designed ontology. Adding context to the information in this way enables more specific questions to be answered through querying the structured data. This allows the system to support a broader range of questions, and produce higher quality answers.

[1] `https://re.ukri.org/research/ref-impact/`

2 The Structured Data Approach

The information held in the REF2014 impact case studies are mainly in the form of unstructured data – an example is shown in Figure1 with key entities and relationships highlighted. Our approach requires this to be turned into a structured form. To achieve this, we designed an ontology that captures the key classes of entities found in the studies and the relationship between them. The ontology design was based on analysis of all the case studies in one of the REF Units of Assessment (UoA11 - Computer Science and Informatics). We also took into account a set of questions that it was important for the system to answer, given our primary interest in how impact was being generated from research. These covered: routes to impact, the quantifiable outcomes of impact, and the underpinning foundations that led to impact (e.g. *publications* and *grants*). We also took into account the need to answer questions found in previous analyses of the unstructured case study text Grant (2015).



1. Summary of the impact

New computational analysis methods have been developed to make drug discovery and toxicological analysis much more efficient. These methods have been patented (UK, EU, US) and are employed in e-Therapeutics Plc, a computational drug discovery spin-off company of the University. The company, introduced to the Alternative Investment Market of the London Stock Exchange in 2007, is now the eighth largest company (by market capitalisation - £92.7M (26/6/2013)) in the pharma/biotech sector. The underlying technologies derive from network analysis and workflow research at the University. The company has an anti-cancer drug (ETS2101) in phase I clinical trials in the UK and the US, and an anti-depression drug (ETS6103) planned to enter phase IIb clinical trial shortly. The beneficiaries of this research are e-Therapeutics directly, other drug companies, and ultimately patients.

Figure 1: Example of an Impact Case Study's Summary of Impact Section

To structure and query the data, we adopted the knowledge representation technology of the semantic web. This includes ways of representing knowledge as Subject-Predicate-Object triples (represented as RDF Beckett and McBride (2004)) conforming to a controlled vocabulary according to an ontology specified in OWL2 Hitzler et al. (2012).The top-level taxonomy is taken from schema.org[2] - a collaborative community that develops schemas for structured data on the internet. We reused their vocabulary and hierarchies to name the classes, sub-classes and relations. However, these classes are not fine-grained enough for our needs. Specific events including "spin-off" and "collaboration" must be represented. We took the categories from Ressearchfish Researchfish (2019) as the starting point for this. Reusing these existing categories not only helped us to focus on designing the relations within the controlled vocabulary but also reduced the ambiguity when introducing a large number of new terms. Designing such an ontology also requires the accurate translation of the unstructured information into structured knowledge. Our first attempt used only **ObjectProperty** to describe the connected entities, but this RDF's inability to attach properties to a relation led to a loss of information. To overcome this we combined the design of the "Action" class from schema.org and categories from Researchfish. This enabled us to use entities to represent rich information about activities. It resulted in a concise and legible 'Activity" class, which nevertheless offers the required level of detail for question answering and further analysis. Part of the hierarchy of the "Activity" class is:

- ***Activity***
  - **Schema:ActivateAction**
  - **Schema:ConsumeAction**
    - UseActivity
      - Public Use
      - Academic Use
      - Industrial Use
    - WatchAction
  - **Schema:CreateAction**
    - Drug Development
    - Job Creation
    - Design
    - Software Development
  - **Schema:InteractAction**
    - Collaborations & Partnerships
      - Collaborative Software Development
      - Knowledge Transfer Partnership
      - Project
    - Engagement Activities
      - Presentation
      - Workshop
    - UnderpinningResearch
    - Influence on Policy Practice Patients and the Public
  - **Schema:TradeAction**
    - BuyAction
    - SellAction
    - Alternative Investment Market Listing
  - **Schema:TransferAction**
    - Awards and Recognition
      - Attracted Visiting Staff
      - Awards
      - Open Competition Winning
    - Funding
      - Further Funding
    - Intellectual Property & Licensing
      - Patent Application Published
      - Patent Granted
    - Investment
    - Spin-off
  - **Schema:UpdateAction**
    - Improvement
    - Service Improvement
    - Technical Improvement
  - ....

This illustrate some of the ways in which impact activities are categorised.

[2] https://schema.org

The full Research Impact ontology shown in Figure 2. As well as *Activites*, the key entities that are modelled include *Collaborations*, *Funding*, *Knowledge Transfer types*, *Patents*, *Grants* and *Papers*.

All of the 248 case studies in UoA11 were manually transferred into the semantic web representation as triples conforming to the ontology. An example of one impact case study structured in this way is shown in Figure 3 (Unit of Assessment 11, Impact Case Study number 21791).

Capturing all the case studies in this form allows us to answer questions using the SPARQL query language Harris et al. (2013). For example, *"Which case studies were based on research funded by the EPSRC?"* is answered by the query:

```
SELECT  ?caseID ?Amount ?Project Leader ?Reference ?Start Date ?End Date
WHERE{
        ?case rdf:type <http://ref2014.ontology/CaseStudies> .
        ?case <http://ref2014.ontology/hasFunding> ?Funding .
        <http://ref2014.ontology/EPSRC>
            <http://ref2014.ontology/COP#fundTo> ?Funding .
        ?Funding <http://ref2014.ontology/Funding:hasAmount> ?Amount .
        ?Funding <http://ref2014.ontology/projectLeader> ?Project Leader.
        ?Funding <http://ref2014.ontology/Funding:Reference> ?Reference .
        ?Funding <http://ref2014.ontology/Timeframe:from> ?Start time .
        ?Funding <http://ref2014.ontology/Timeframe:to> ?End time
}
```

It produces the answer shown in Table 1.

We can also answer questions on relationships, such as *"In what ways were IBM involved in the case studies?"*:

```
SELECT ?Company ?Activity ?Object
{
        ?Company rdf:type <http://ref2014.ontology/Company> .
        ?Company rdfs:label "IBM"@en .
        ?Company ?relation ?Activity .
        ?Activity rdf:type <http://ref2014.ontology/Activity> .
    ?Activity ?relation2 ?Object .
}
```

It produces the answer shown in Table 2.

Company	Activity	Relation With
IBM	Acquisition28081	Transitive
IBM	Collaboration12998	Software Migrations Ltd
IBM	Collaboration1672	University of East Anglia
IBM	Collaboration21273	Newcastle University
IBM	Collaboration5801	Swansea University
IBM	IndustrialUse43323	TrOWL
IBM	ServiceProvision21273	WebSphere
...	...	...

Table 2: The Involvement of IBM in the Case Studies.

We can also exploit the structure to aggregate information from across the set of case studies, e.g. *"What was the total value of EPSRC funding that supported the impact case studies?"* The query to answer this question is:

```
SELECT  (SUM(?Amount) as ?amountSum) ?Currency
WHERE{
        ?case rdf:type <http://ref2014.ontology/CaseStudies> .
        ?case <http://ref2014.ontology/hasFunding> ?Funding .
        <http://ref2014.ontology/EPSRC>
            <http://ref2014.ontology/COP#fundTo> ?Funding .
        ?Funding <http://ref2014.ontology/Funding:hasAmount> ?Amount .
        ?Funding <http://ref2014.ontology/Funding:Currency> ?Currency.
}
```

This returns £200,258,937.

2.1 Integrating External Data Sources

One of the advantages of adopting the structured approach is that it enables the knowledge base we have created for the case studies to include links to other relevant, external data sources. For example, where a company appears in an impact case study, we create a link to its entry in DBpedia[3] and to its Company House records[4] to provide additional information, including the Standard Industrial Classification Code. This enables us to support queries that seamlessly combine information from the impact case studies and external data sources. An example is that we can answer the question: *"In which countries are the companies included in impact case studies based?"* The query to answer this question (generating the total for each country) is:

```
SELECT ?Country (COUNT(?Country) AS ?Count)
WHERE{
        ?Company rdf:type <http://dbpedia.org/ontology/Company> .
    OPTIONAL{?Company <http://dbpedia.org/ontology/locationCountry>
?Country.}
    SERVICE <http://dbpedia.org/sparql>{
    ?Company <http://dbpedia.org/ontology/locationCountry> ?Country.}
}GROUP BY ?Country
ORDER BY DESC (?Count)
```

It produces the answer in Table 3:

Country	Count
United Kingdom	312
United States	195
Germany	15
France	13
Japan	8
China	6
Spain	5
Belgium	5
Italy	4
Netherlands	4
Australia	3
Canada	3
Finland	3
India	3
Sweden	2
Switzerland	2
...	...

Table 3: The result of a federated query that uses DBpedia to determine where companies included in the case studies are based.

[3] https://www.dbpedia-spotlight.org/api

[4] https://beta.companieshouse.gov.uk

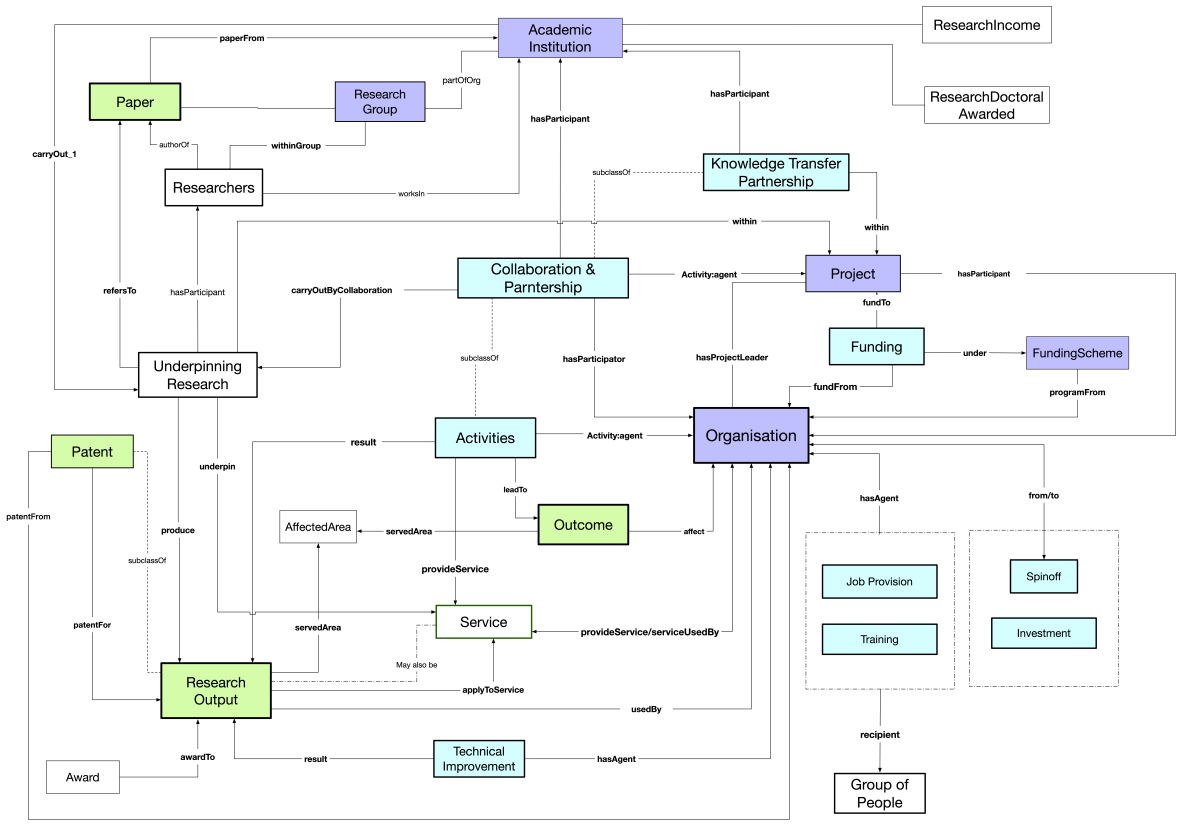


Figure 2: The Research Impact Ontology. Purple nodes relate to Organisations, Cyan to Activities, Green to Research Outputs.

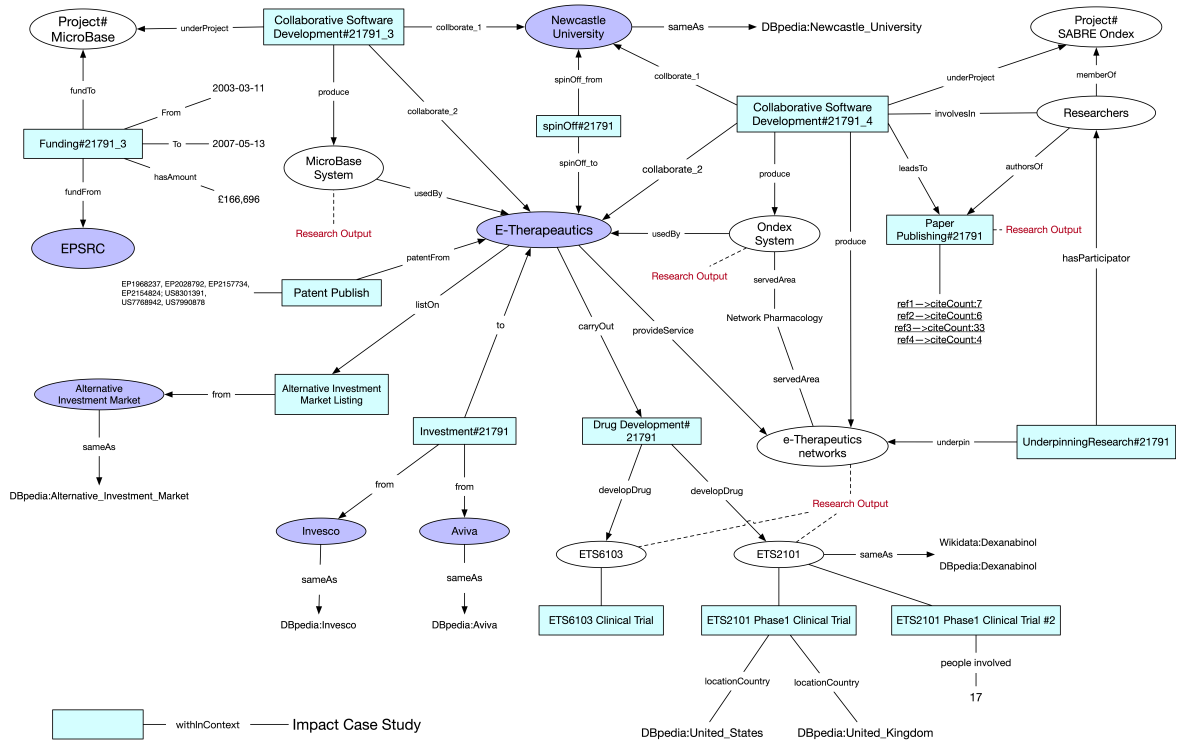


Figure 3: An example of the graph representation of one Impact Case Study

CaseId	Amount	Project Leader	Reference	Start Date	End Date
id12531	£63,429	Nottingham Trent University	GR/R32468/01	01 Mar 2001	28 Feb 2002
id13783	£7,566	Brunel University London	EP/E055141/1	31 Mar 2008	30 Mar 2009
id5800	£5,820,840	University College London	EP/G059063/1	01 Oct 2009	31 Jan 2016
id42146	£6,119,249	Imperial College London	EP/H009744/1	01 Oct 2009	31 Mar 2015
id13830	£283,680	University of Cambridge	GR/S01894/01	01 Jan 2003	31 Dec 2005
id44159	£688,578	University of Southampton	GR/T10664/01	25 Feb 2005	24 Mar 2010
id35118	£887,750	Newcastle University	EP/K006568/1	05 Feb 2013	31 Dec 2016
id2010	£391,850	University of Kent	EP/E049419/1	01 Oct 2007	31 Mar 2012
...	...	...	...	...	...

Table 1: Impact Case Study Research funded by EPSRC.

We have integrated other sources of external information to increase the scope of the queries the system can answer. These include linking company entities to the Company House pages[5] to provide entity properties like Standard Industrial Classification Code and company types for *Company*. We also use web crawlers to retrieve citation counts of *Publications* during the REF2014 period, the h-index of *Researchers*, amount of *Investment*, number of *People* affected during a specific type of *Activity*, etc.

3 Summary

The paper describes how creating a structured knowledge base makes it possible for policymakers to answer a broader range of questions about evidence than is possible with free text data. This is because the structured representation gives a context to the terms included in the text. We have also shown that it is possible to link to external data sources, so supporting questions that rely on this additional information. The focus of our work has been on supporting those seeking to extract value from the REF2014 Impact Case Studies, and a significant part of our work was designing the ontology to structure this data. However, the overall approach is applicable to other areas where policymakers wish to analyse evidence. We are now using the structured data as the basis for applying machine learning to extract interesting patterns from the data, and to see if the grades awarded to the case studies can be accurately predicted.

References

Beckett, D. and McBride, B. (2004). Rdf/xml syntax specification (revised). *W3C recommendation*, 10(2.3).

Grant, J. (2015). The nature, scale and beneficiaries of research impact: An initial analysis of research excellence framework (ref) 2014 impact case studies. Technical report, King's College London and Digital Science.

Harris, S., Seaborne, A., and Prud'hommeaux, E. (2013). Sparql 1.1 query language. *W3C recommendation*, 21(10):778.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., Rudolph, S., et al. (2012). Owl 2 web ontology language primer. *W3C recommendation*.

Researchfish (2019). Research outcomes common question set. Technical report, Researchfish.

[5] `https://beta.companieshouse.gov.uk`