The Use of Open Data to Improve the Repeatability of Adaptivity and Personalisation Experiment

Harshvardhan Pandit, Roghaiyeh Gachpaz Hamed, Shay Lawless, David Lewis
ADAPT Centre, Trinity College Dublin, Ireland
Emails(@adaptcentre.ie): harshvardhan.pandit, ramisa.hamed, seamus.lawless, david.lewis

ABSTRACT

Reproducibility of results is a key element for the verification of scientific experiments and an important indicator of the quality of a published experiment. It is vital therefore to precisely and transparently share both the method and the data associated with an experiment. Data associated with an experiment is often linked within peer-reviewed scientific publications, and is difficult to assess in a consistent manner. In this paper we explore how emerging linked data standards can be applied to the description and data of published adaptivity and personalisation experiments in a manner that can be linked from publications and easily located, accessed and reused to repeat an experiment. The approach also provides possibilities for published experiments to be extended or modified to provide a firmer grounding for publishing new results and conclusions.

Keywords

Linked Open Data, Language Resource Provenance, Natural Language Process Flows

1. INTRODUCTION

Compared to other sciences, experiments in adaptive software are the technological aspects of implementation, i.e. the software implementing different parts of the experiment. The publication of source code, especially in open version control systems such as GitHub[1] renders the version of a software component equally referable from a publication as well as the data sets involved. Repeatability of experiments therefore relies on the known rights to reuse both the code and the data. While usage licenses for source code are well established, the usage rights for experimental data can still present obstacles, especially when, as is common in personalisation experiments, it contains specific differentiating data related to individual experimental subjects.

As the Adaptivity and Personalisation scientific community considers a more structured approach to comparative experimentation, we examine how this can leverage the state of the art in both Linked Open Data and open science best practices to establish a cutting edge infrastructure for repeatability in experimentation. An influence is the Natural Language Processing (NLP) community, which has a well established practice of shared tasks and competitions. This is especially relevant as an exemplar of scientific community best practice because NLP and Machine Learning (ML) components operating over structured and unstructured data are becoming increasingly important as parts of research in the UMAP community. Also important is the work in establishing open meta-data for scientific data.

2. LINKED OPEN DATA

Linked Data is the interlinking of structured data through semantic queries using web technologies. This is distinct from the use of the term in some disciplines that use linked data to describe commonality of data between sources.

Linked Open Data is based upon the principle[2] of interlinking resources and data with RDF and URIs that can be read and queried by machines through powerful querying mechanisms like SPARQL. Adaptive and personalisation experiments can benefit from declaration of information such as usage rights, provenance, and authorship to create a more declarative and open approach for sharing research knowledge and experimental data. Ontologies such as DCAT[2] help in expressing authorship of workflows and data sets, whereas ODRL[3] expresses usage rights and licensing. The provenance of an experiment can be captured and modelled using the PROV[4] family of ontologies. The use of distinct ontologies for experiment steps and data allows the publication and discovery of experimental data with specific metadata such as usage rights that can be collected or aggregated to form linked repositories such as OpenAire[5] and Linghub[6]. Existing research for these standard vocabularies has examples of best practice[1] for publishing data sets' metadata as linked open data. The machine readable nature of metadata makes it easy for an automated system to verify the correctness of the data, or perform other operations that may prove helpful in validations such as checking of data formats and

[1]https://github.com

[2]http://www.w3.org/TR/vocab-dcat/
[3]http://www.w3.org/TR/odrl/
[4]http://www.w3.org/TR/prov-0/
[5]https://www.openaire.eu
[6]http://linghub.lider-project.eu

alterations through replacement of key steps, which provides opportunities for new outcomes and results.

Repetition and variation form the bulk of research in Natural Language Processing (NLP) and Machine Learning (ML) experiments, with several community-led projects that aim to share experiments as a set of metadata using linked open data vocabularies. A broadly accepted workflow modelling process provides the opportunity to combine various data sets into a collective corpus that increases the range of available annotated data. This can be leveraged for the discovery of experiments and data sets based on their metadata through aggregated repositories such as LingHub. By linking experiments as resources with the data used or produced, variations in experiments can be evaluated using existing data sets for a better comparative evaluation.

The structured sharing of an experiment through metadata allows others to build upon the experiment by modifying its steps or reusing its components. The process of providing an experiment as a linked resource reduces duplicity in research and allows promotes reuse of knowledge and repeatability. By approaching a definitive workflow framework as in the ML and NLP community, a system of expressive models could provide useful insights and design considerations for personalisation and user modelling systems. Such an approach would leverage previous research in workflow modelling such as MEX[3] and NIF. The NLP community has developed a schema, termed META-SHARE[5], for language resource metadata that shares many characteristics with the OpenAire metadata scheme. The META-SHARE schema has also been mapped into RDF with relevant attributes mapped to specific properties from the above standard vocabularies and is used by LingHub as an aggregation source.

3. EXPLORING NEW TECHNOLOGIES IN DATA FORMATS AND ONTOLOGIES

An experiment method is often comprised of several steps that are connected by the data exchanged between steps. By providing information about the steps along with the data being used, it is possible to repeat or verify certain steps without repeating the entire experiment. The separation of steps also enables replacement of a step with other comparable approaches and to evaluate comparative results between them. Such a workflow lends itself to ease variation in experiment repeatability and makes it possible to compare results across a range of similar experiments. The abstraction of experiment workflows from individual steps also allows each step to be implemented using different technologies. The information or metadata about the experiment and each step can be expressed efficiently using the P-Plan[7] ontology that expands upon PROV for representing execution workflows.

When dealing with adaptive and personalisation systems, data forms an important set of resource for comparative evaluations. The practical implementation for such systems is sometimes designed based on performance or viability to existing practical considerations. CSV on the Web (CSV-W)[8] is an adaptation of the widely used CSV format for linked data sets that allows representing structured data along with a metadata vocabulary for describing the contents of the data. This makes it possible for a system to be performant while using the CSV-W format for all forms of data exchange including the metadata and the actual data set related to an experiment. Similarly, JSON-LD[9] is based on the JSON format, which is a popular format for exchanging data in a structured manner across the various REST APIs across the web. JSON-LD is a lightweight data format that is easy for humans to read and write, and is currently a W3C recommendation. CSV-W and JSON-LD offer a fast and performant way to exchange RDF/OWL based data in linked open data systems.

[7]http://www.opmw.org/model/p-plan/

[8]http://www.w3.org/TR/csvw-ucr/

4. EXAMPLE USE CASE

We applied the combination of DCAT, PROV, P-PLAN and CSVW metadata to tabular data sets arising from an experiment using the Personalized Multilingual Information Retrieval (PMIR) platform described in[4]. This platform supported the modular design, implementation and evaluation of a set of algorithms for multilingual user modelling, multilingual query adaptation, and multilingual result-list adaptation. A simplified summary of the user modelling process is expressed in Figure 1. It shows the structure of the workflow, including steps and variables involved and their dependencies on state and execution order captured using the P-Plan vocabulary. Steps are defined with additional metadata describing dependency relations and precedence order in relation to other steps along with the data consumed or generated, which is expressed as variables. The use of an expressive ontology provides an interoperable and unambiguous record of the workflow interlinked with the resulting data set.

5. FUTURE WORK & MOTIVATION

The creation of an ontology specific to the description of the steps and data associated with a published adaptivity and personalisation experiment offers a cohesive model that can be linked together to form a corpus of knowledge of related experiments. The efforts required to create such an ontology are heavily based on adopting existing ontologies and standards and have the advantage of best practices for models that easily locate, access, reuse and repeat an experiment. Sharing experiment workflows and experimental data allows broader sharing of knowledge linked across domains. This brings new ideas into the UMAP community while exposing the ongoing work in new research avenues to other fields of related disciplines. Ultimately, scientific research can only progress through sharing of knowledge in a usable and repeatable manner and hence must endure efforts towards the same.

Acknowledgements

This work has been supported partially by the European Commission as part of the FALCON project (contact number 610879) and the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

6. REFERENCES

[1] M. Brümmer, C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, and S. Hellmann. Dataid: Towards

[9]http://json-ld.org

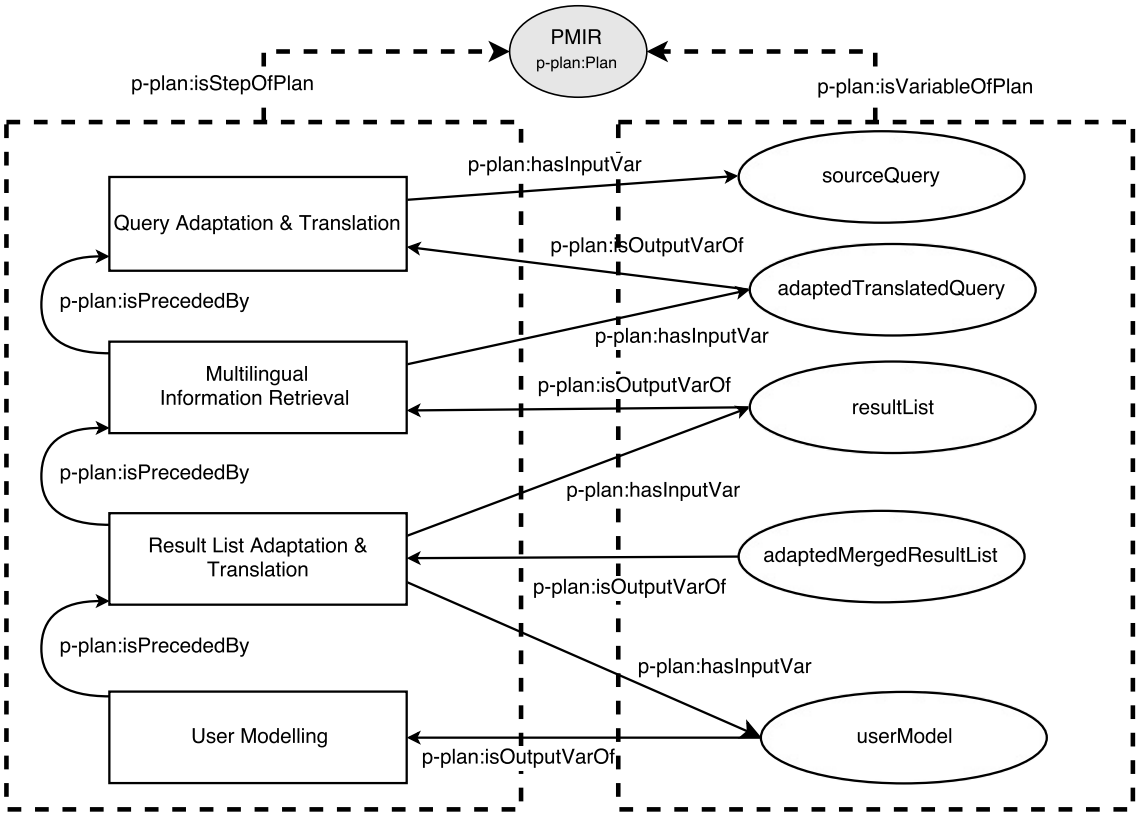

Figure 1: Summary of P-Plan model for Experiment

semantically rich metadata for complex datasets. In *Proceedings of the 10th International Conference on Semantic Systems*, SEM '14, pages 84–91, New York, NY, USA, 2014. ACM.

[2] T. B.-L. Christian Bizer, Tom Heath. Linked data - the story so far. In *Special Issue on Linked Data*, pages 5(3) 1–22. International Journal on Semantic Web and Information Systems, 2009.

[3] D. Esteves, D. Moussallem, C. B. Neto, T. Soru, R. Usbeck, M. Ackermann, and J. Lehmann. MEX vocabulary: a lightweight interchange format for machine learning experiments. pages 169–176. ACM Press, 2015.

[4] M. R. Ghorab, S. Lawless, A. O'Connor, and V. Wade. Does personalization benefit everyone in the same way? multilingual search personalization for english vs. non-english users. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014.*, 2014.

[5] S. Piperidis. The meta-share language resources sharing infrastructure: Principles, challenges, solutions. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *LREC*, pages 36–42. European Language Resources Association (ELRA), 2012.