Engaging Content
Engaging People

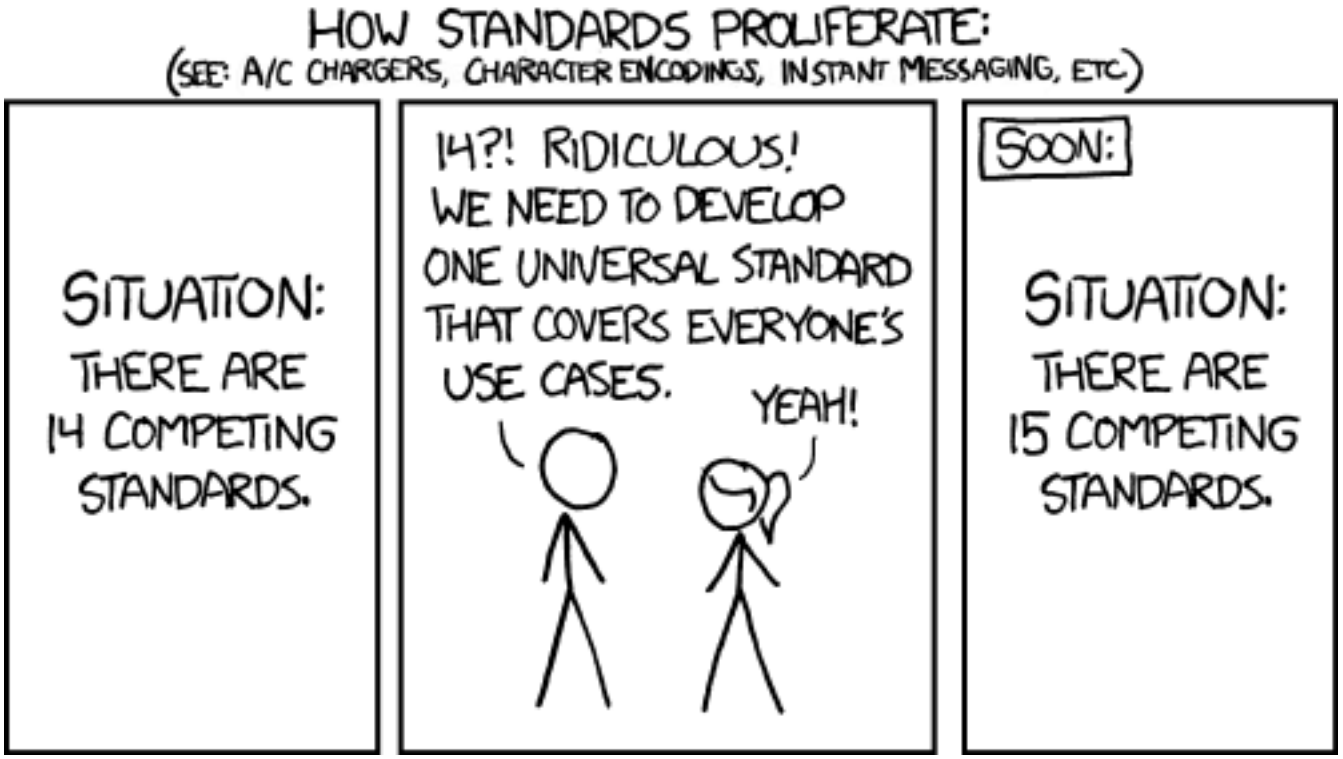
The Use of Open Data to Improve the Repeatability of Adaptivity and Personalisation Experiments

Harshvardhan Pandit, Roghaiyeh Gachpaz Hamed,
Séamus Lawless, David Lewis

ADAPT, Trinity College Dublin
Dublin, Ireland

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
SFI Science Foundation Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Reproducibility

Reproducibility of results is a key element for the verification of scientific experiments and an important indicator of the quality of a published experiment.

- Increasing emphasis on open access to data supporting publicly funded research, e.g. EU H2020
- Open Access Repositories emerging, e.g. OpenAIRE
- Open vocabularies for data in different research domains, e.g. Research Data Alliance

Linked Open Data

- Linked data standards from the W3C can be applied to the description and data of published experiments
- Standard data vocabularies provide machine-readable meta-data for indexing and searching across repositories
- Can be linked from publications and easily located, accessed and reused to repeat an experiment.
- Published experimental data can be extended or modified to provide a firmer grounding for publishing new results and conclusions.

Usage license and rights

- Experimental data may not always be freely available
  - e.g. data from industry collaborators
- Who can use it? Under what conditions?
- Is the usage license of the data different and separate from the usage license of the experiment?
- Personalisation experiment data that identifies individuals cannot be openly published
- Some of it may be anonymised, but how effectively? Does this impact reproducibility?

Natural Language Processing and Machine Learning

- Highly structured, Highly repeatable experimental culture in Language technologies
- Most research based on adaptation and modification of existing experiments using well defined metrics - leads to rapid pace of experimentation
- Experimental data often defined specifically for reuse as part of Shared tasks and Competitions

Scientific Applications of Linked Data

- LingHub aggregated and offer search services over >100,000 language resources meta-data from different repositories using linked data principles
- NLP and ML benefit from open data formats that ease exchange between platforms
    - NLP data Interchange Format (NIF) and the Machine Learning Experimental Vocabulary (MEX) from University of Leipzig

Integrate Existing LOD Vocabularies

- RDF / OWL:
  - Knowledge representation
- Data Catalogue (DCAT):
  - Authorship and Publication meta-data to help cataloguing data
  - Declare authors and contributors of experiment
- Open Digital Rights Language (ODRL):
  - Digital Rights and Licensing
  - Attach terms of usage, publication etc. to data or experiment
- PROV / P-Plan:
  - Provenance: logs entities, activities performed on them and agents performing activities
  - P-Plan captures data flow between activity steps
- Open Provenance Model for Workflows
  - For Scientific Workflow
  - Extends PROV / P-Plan to define scientific workflow templates and link to data set from individual workflow execution

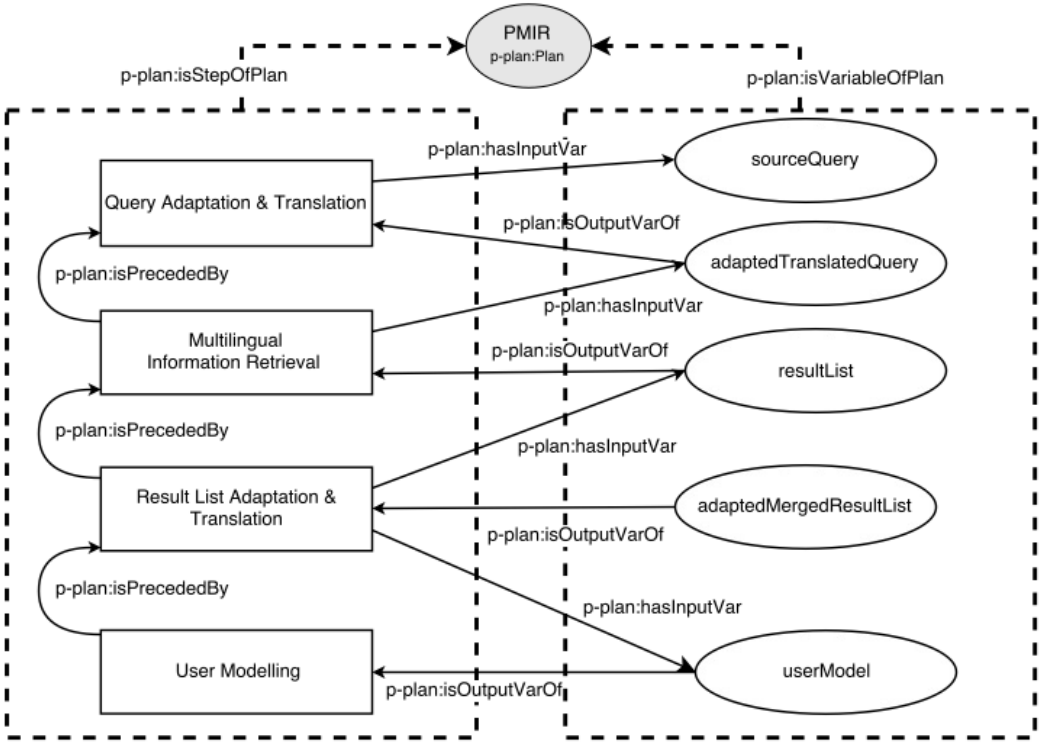
Technology – making LOD easier

- CSVW (CSV for the Web)
  - structured linked data using CSV
  - CSV is widely used to share data sets
  - CSV is a format that is widely supported
- JSON-LD (JSON for Linking Data)
  - RDF metadata representation using JSON
  - JSON is lightweight and popular, easy to integrate into Web applications
  - share metadata for experiments

Experimental Data - Use Cases

- An experiment can have multiple steps, and data can be exchanged between steps.
- It should be possible to repeat or verify certain steps without repeating the entire experiment.
- You can then replace a step with other comparable approaches and evaluate comparative results between them.
- It is then possible to compare results across a range of similar experiments.
- The abstraction of experimental workflows from individual steps also allows each step to be implemented using different technologies.

Personalised Multilingual Information Retrieval (PMIR)



LOD for UMAP Evaluation Community

- Existing LOD vocabularies provide the basis for publishing and sharing experimental data widely via the Web
- Can capture workflow templates and logs, licensing declarations, cataloguing meta-data
- Extensive tooling available for storage, queries, publishing, APIs and unit tests
- Easily accessible using familiar web application development technologies
- Aligns well with public policy and community trends in open scientific data
- Needs specialisation of workflow vocabulary for user modelling and personalisation experiments
    - e.g. classifying common experimental steps and data variable types
- Currently developing tooling specifically for publishing, reusing and branching experimental Language Technology workflows, including shared task support

Acknowledgements

Thank you!

Questions? Comments?

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.