In the Pursuit of Open Science

Alexander Garcia[1,*], Leyla García[2], Olga Giraldo[3],
[1]Institute for Digital Information and Scientific Communication, Florida State University, Tallahassee, Florida, USA. Temporal
[2]Knowledge Bases Group, Universitat Jaumé I, Castello de la Plana, Valencia, Spain. [3]Ontology Engineering Group. Facultad de Informática, Universidad Politécnica de Madrid, Spain.

In the Pursuit of Openness

"*Life, Liberty, and the pursuit of Happiness*" is a paramount phrase in the United_States Declaration of Independence. The phrase illustrates the various "unalienable rights" which the Declaration says all human beings have been given by their Creator and for the protection of which they institute governments. Although “we the research community” enjoy various “unalienable rights" inherent to our work, the right of controlling our creations is not one of them. Moreover, in a similar way to that of the early colonies inheriting and adapting an onerous or tyrannical form of government, “we the researchers” are seeking more fairness and justice for the outcomes of our work –the content we generate is as diverse as research itself.

Citations, commonly referred to as citation data, illustrate an interesting situation arising from old practices in new times. Although authors carefully gather references supporting statements, once the publication is available as a PDF, citation data is significantly locked up. Furthermore, basic bibliographic metadata pertaining to the PDF is not always available. Meaningful entities such as figures, tables, captions, domain specific names and author information are often lost for practical purposes. Extracting information from PDFs as well as embedding content from PDFs into the Web of Data is challenging, though possible, but is an unnecessary barrier. How should authors link specific parts in documents to datasets? Once content is machine processable, how do we best support social participation around particular parts of the document? Science is becoming less dependent on one final document and a single central narrative; it is heading towards an aggregation of context and problem dependent self-describing research objects. Such a fluid array of research objects should support shareability, reproducibility, discoverability and reusability. Although interoperability is central to open science, “we the people” are struggling in the face of change.

Scientific content is varied, takes shapes known as scientific papers, database entries, presentations, code, etc. Although diverse, scientific content is usually related to a specific research question, hypothesis, or problem statement. By the same token, these are usually brought together in research projects. Research objects are therefore difficult to define and constrain.

Here, we present our approach to research objects as semantic aggregators. Central to our approach is the interoperability across structured ROs. We argue that ROs should be self-describing artifacts living in a fluid grid of relationships, a hyperontology that provides dynamic structuring mechanisms. Within our approach, documents, as self-describing entities, are not central to the communication of research; there is, however, a narrative structure that should be automatically built depending on the aggregation and structuring of ROs. Such aggregation is context and problem dependent; documents within this scenario are to be born semantic. Also fundamental to our approach is the idea of documents and research related files being semantic from their conception.

We illustrate our approach with two scenarios. The recreation of the research reported by Gutierrez et al in “Identification of a Rice stripe necrosis virus resistance locus and yield component QTLs using Oryza sativa × O. glaberrima introgression lines”. Within this scenario, we present ROs, ancillary research-related files, laboratory notebooks, LIMS records, laboratory protocols, results, code, etc. We semantically structure all entities within a hyperontology; we then present a simple dynamic context dependent aggregation over such fluid array.

ACKNOWLEDGMENTS

Special thanks to Robert Stevens for his comments on the manuscript.

References

[1] Andrés G Gutiérrez, Silvio J Carabalí, Olga X Giraldo, César P Martínez, Fernando Correa, Gustavo Prado, Joe Tohme andMathias Lorieux. Identification of a Rice stripe necrosis virus resistance locus and yield component QTLs using Oryza sativa × O. glaberrima introgression lines. In BMC Plant Biology. 2010.