Linked Data and Text Mining as an Enabler for Reproducible Research

John P. McCrae, Georgeta Bordea and Paul Buitelaar
Insight Centre for Data Analytics, National University of Ireland, Galway
{john.mccrae, georgeta.bordea, paul.buitelaar}@insight-centre.org

Abstract

Research data is one of the most important outcomes of many research projects and a key for enabling reproducibility in the analytic data sciences. In this paper, we explain three main challenges that complicate reproducibility namely, the difficulty of identifying datasets unambiguously, the lack of open repositories for scientific data and finally the lack of tools for understanding published science. We consider the use of linked data and text mining as two tools to solve these issues and discuss how they may ameliorate these issues.

Keywords: Reproducibility, data science, metadata, linked data, text mining

1. Introduction

Research data is increasingly becoming not only an important outcome of any research, but also often the key to ensuring that this research is reproducible, as most scientific experiments consist partly or entirely of data analysis (Borgman, 2012). Thus, analytic reproducibility is a key validation of a scientific result and so it is vital that researchers have access to the data for experiments and the code and processes used to perform these experiments, yet it has been identified that the management of research data is currently quite insufficient (Piwowar and Chapman, 2010). In fact, the reality is that most research data is made available only after the research project has been completed and then often becomes unavailable within only a few years of the end of the project. Even worse, the quality of these datasets often falls short of even basic standards (Kontokostas et al., 2014), even though best practices from software engineering have shown that principles such as continuous testing and version management should be considered from the start of the project. A fundamental challenge that needs to solved is the ability to identify, discover and describe research data and thus for datasets to be federated between trusted repositories and discoverable by means of persistent identifiers and metadata. It is our experience that researchers are in fact very willing to create data of high quality, but they are not supported by the right tools, and moreover they certainly do contribute data when journals make depositing open data a requirement for submission (Wellcome Trust, 1997; GenomeCanada, 2005). The use of linked data, semantics and natural language processing techniques can be combined to make a researcher-friendly architecture that allows high quality research data and analytic reproducibility of research results to become the norm.

In order to meet this goal we believe that a combination of techniques built around open networks is necessary. As such we propose three main components that we believe are essential for ensuring reproducibility in analytic data sciences, by which is meant experiments that are based on analysis of data by means of various algorithms. Firstly, we need a system that can unambiguously identify the data and all processes applied to the data. Secondly, we need a system that can provide the descriptions of these experiments and provide means to look-up systems and experiments and enable them to be executed in an 'on-demand' fashion. Finally, we recognize that the effort of creating sophisticated metadata is largely too onerous for the typical researcher and creates very little value for her or him. As such, we propose that we build on existing text mining technologies to extract the necessary descriptions from published scientific papers and practical descriptions. This will also allow us to retroactively include the large amount of research already done into repositories of such information.

2. Identifying Research Data

The most typical method currently used for identifying research data is by means of (HTTP) URLs and this has some advantages, most notably that it is clear to all users how the resource can be located and it includes some information about the provider (or at least maintainer) of the resource in the form of the domain name given in the URL. However, a crucial weakness of this schema is that HTTP URLs identify a particular file on a single server and many things from failure of service, departure of managing personnel or simply neglect at the end of a project can cause this server and/or file to become unavailable. As such the current practice of quoting HTTP URLs in research outputs in practice discourages reproducibility in research.

An alternative option is to define a fixed identifier that identifies the resource, such as Digital Object Identifiers (Paskin, 2008, DOI), or in the particular community of language resources the International Standard Language Resource Numbers (Choukri et al., 2012, ISLRN). These systems have had less success than URLs in research. One of the reasons for this may be that they are unstable in that they are owned by a particular organization or group of organizations and depend on the continuous maintenance by these organizations. While it seems unlikely that the coalitions behind these schemes will dissolve soon, on the scale of 50 to 100 years technology changes may make this a high likelihood. More likely, the primary reason for the lack of adoption of these systems is that they provide a significant barrier to entry with many researchers being simply unclear about how to assign a value to a resource. In particular, such schemes may prove to be difficult for so-called 'citizen scientists' (Cohn, 2008), who contribute data by crowd-sourcing alongside professional scientists.

In order to provide true digital preservation, the principle of 'lots of copies keep stuff safe' pioneered in the eponymous LOCKSS system (Maniatis et al., 2005), seems vital. How-

Figure 1: An example of double hashing

ever, this network relies on a complex voting procedure to ensure stability and has thus only be installed principally by university libraries.

The use of an algorithmic identifier such as secure hash[1] to identify the dataset would be an interesting option in this situation. However it has several clear disadvantages: firstly, of course there is a risk of collision, i.e., two hash codes may have the same value. This can easily be mitigated by using codes of a certain length, for example a 72-bit code can be easily represented in 12 Base64 digits[2] and the mathematical expectancy of the first collision is only after 100 billion objects have been identified. Further, assuming that these codes can be easily resolved, a simple check for duplicates should allow collisions to be avoided. Other issues are that such a schema does not include any identification of the authors and as such it may make more sense to perform a *double hash* (illustrated in figure 1), that is, first hash the dataset, then include the hash in a standardized metadata format and calculate the hash of the metadata document. One of the major advantages of this scheme is that once published a dataset cannot be changed, thus ensuring that the resource described in a paper is exactly the resource used in the authors' experiments. Another advantage of this is that it is easy to add an extra nonce parameter in the unlikely event of a hash collision.

An important aspect of reusing any data is the metadata and documentation that goes along with this. This metadata and documentation is likely dynamic and will change and be updated, and for this reason it makes sense to build this metadata as linked data so that it is possible to take advantage of the links to provide more information and allow the data to be self-documenting and packaged as research objects.(Bechhofer et al., 2013) However, some metadata is necessary to enable re-use of the dataset including the license of the dataset, links to documentation, basic description and citation information. As this part of the data is static, we propose that this basic metadata profile is provided in the metadata document that is hashed in this scheme and that his metadata is expressed using RDF.

The combination of linked-data-based metadata with double hashing provides a powerful option for the creation of linked data repositories whereby the data can be described using open, flexible metadata parameters, that are further refined with semantics on the Web. These metadata descriptions could easily be converted with this scheme, allowing multiple heterogeneous repositories to share and integrate resources based on a single identifier (based on double-hashing) and a single underlying format (RDF) that would allow data to be stored and shared for the entire lifetime of the dataset, not just the project that created it.

It is of course, an issue that citations may only refer to parts of a dataset and as such the use of identifiers to identify parts of the dataset, for example the Media Fragment URIs (Troncy et al., 2012) or RFC 5147 (Wilde and Duerst, 2008) should also be employed.

3. Repositories for Analytic Data Science

One of the key issues with research data is that it is currently very poorly and heterogeneously described, which acts as a significant barrier to access. In a recent analysis (McCrae et al., 2015) it was shown that among four major collections of information about language resources only 5.2% of resources appeared to be contained in more than one repository. Moreover, we found that even basic metadata properties had a large disagreement about how they were to be represented, e.g., a language may be represented by its English name, or using one of the ISO codes, and that key properties about the resource, such as its license were missing in most cases, e.g., only 3.0% of metadata records gave the description of the resource.

As such, it is clear that most centralized approaches to metadata collection are insufficient and we need to develop systems that can aggregate and improve data. Such a system would need to integrate heterogeneous data sources and provide links to each of the sources and the original datasets. It seems natural that linked data (Bizer et al., 2009) would be helpful here as it allows for metadata that is heterogeneous, extensible and easily aggregated from multiple sources. It is our principle belief that many of the tools for creating and using metadata records such as RDF (Klyne and Carroll, 2006), DCAT (Maali et al., 2014) and SPARQL (Prud'Hommeaux et al., 2008) are already in existence, however, none of these are specific to scientific workflows or any specific scientific domain, and key vocabularies for versioning and quality certification are absent. As such, it is vital that we extend existing schemas to provide a more complete description of the data described and how it can be used for scientific reproducibility.

Moreover the entire analytic research program can be recast as research data, either by formal description of processes and workflows or by embedding process in software containers, thus transforming complex analytic experiments into single binary files. There have been a number of systems proposed for modelling scientific workflows such as

[1] Similar to methods employed by the GIT versioning system to identify individual commits

[2] For example: `MC4yMzIzNTU2`

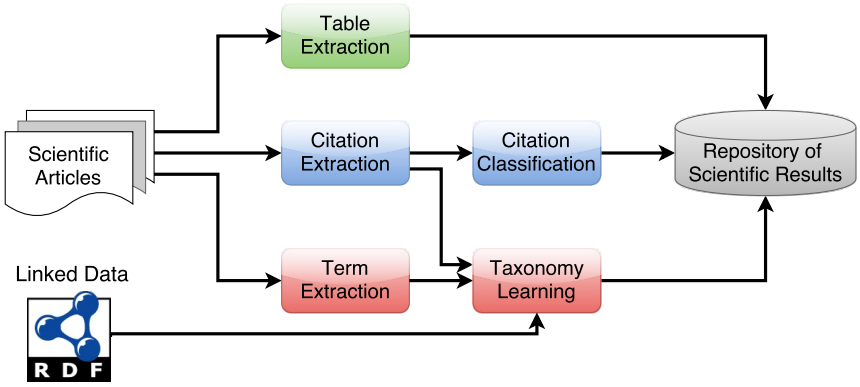

Figure 2: The architecture of a text mining system for analyzing the scientific literature

myExperiment (Goble et al., 2010). These have significant barriers to entry, most notably that work flow engines are difficult to learn, hard to apply and do not truly guarantee the same result every time, as external libraries may change. A much simpler solution is to use virtual machine (VM) images, something that has become much easier and quicker due to recently developed technologies[3], that allow the entire process to easily be stored including the exact state of the whole system used to be described. This approach also reduces adoption costs as the original authors need only install the system once in the VM or software container and then the software can be used on any platform supporting the container software. For single-machine experiments a single VM image and a command could allow for quick and easy 'one-click' reproduction of scientific results. For more complex runs the use of multiple VM images still significantly reduces the process of describing workflows.

Another issue that still needs to be handled in the context of providing research data is that of versioning, in particular tracking the development of resources that have been created in a collaborative manner. A recent development of the Global WordNet Association, the Collaborative Interlingual Index (Bond et al., 2016; Vossen et al., 2016) has shown that for resources that can be quickly updated with minor changes, the use of a version control system[4] can help with the development of the resource and can ensure that a particular version can be cited. As such, the use of version control as a primary part of the scientific work would allow for the metadata about resources to easily be accessed and exported to a queriable interface and such a system is already under development[5].

More importantly, metadata is created by humans in natural language and it is our experience that natural language processing techniques, in particular semantic textual similarity, are required to ensure that descriptions are truly interoperable. This is particularly true if we assume that we will not have direct control over the metadata creation process but instead must ensure harmonization of metadata for external sources is performed post-factum. As such, it is necessary to look into techniques in such as vocabulary alignment (Euzenat et al., 2004) in order to create and consolidate metadata files and novel techniques, including using semantic textual similarity on descriptions (Xu et al., 2015) will further automate this process, however more research is needed in this area.

Thus we require the creation of a platform for the management of data and processes built on existing software engineering methodologies including Git and Docker and continuous integration, whereby the scientific improvements can be clearly visualized and the reproducibility is open and achievable with a single click.

4. Text Mining from the Scientific Literature

In spite of the effectiveness and ease-of-use of any potential system for managing research results it is natural that any system will not achieve complete adoption. Moreover, there is still a large amount of scientific experiments that have already been conducted. For these reasons, it is necessary to analyze the already conducted literature in particular looking to identify:

Data Any datasets used in a research paper as well as the links to these datasets and the version information if available.

Method The methods used in the paper, in the form of the names of algorithms and if possible the links to the code used.

Results What results are reported by the authors and what metrics and methods were used to achieve these results.

This will create a database of basic scientific facts similar to existing proposals such as "Nanopublications" (Groth et al., 2010). A first step to automatically extract information about datasets, algorithms, and results from a scientific publication is to capture the internal structure of the document and to identify relevant sections and paragraphs. Authors use section headings to explicitly mark experimental sections, but there is some variation across research domains and communities. Supervised approaches are particularly well suited for this task.

Scientific articles often provide empirical evidence to support a novel approach by presenting extensive comparisons

[3] In particular, Docker `http://www.docker.com`

[4] In this case Git

[5] `http://conquaire.uni-bielefeld.de/`

with state of the art approaches, using multiple datasets that are either introduced by the authors themselves or that are constructed and made available in related work. The typical way to reference these external algorithms and, in some fields, data sources is by using citations. Therefore, citation extraction and resolution plays an important role in identifying as accurately as possible all the investigated datasets and methods. With several solutions readily available, this is a relatively straightforward step.

Evaluation results are usually provided in tables, therefore the ability to find tables and extract information from them (Pinto et al., 2003) is crucial for extracting this type of information. Because of space constraints, authors liberally make use of acronyms to refer to datasets and algorithms, therefore acronym detection and resolution is also important.

Extracted information about datasets, methods, and results can then be used to populate large repositories about experimental results. But these would largely be unusable without storing as much context as possible about provenance, date, research topics, experts involved and how to contact them. A solution for this could be to build on an existing text mining system, such as Saffron[6] (Monaghan et al., 2010), which currently offers support for keyphrase extraction, entity linking, taxonomy extraction, expertise mining, and document browsing for scientific publications. Currently the system generates automatically constructed taxonomies of scientific topics to support search and discovery of scientific publications, but the system can be easily extended to offer similar support for locating experimental data. An architecture for such a system is shown in Figure 2.

5. Conclusion

It is increasingly true that "science depends on good data" (Whitlock et al., 2010, p. 145) and as such the management of data will become one of the central activities for all scientists and many researchers in the humanities. Currently, much of the response to these challenges has been institutional, in that large networks of institutes and researchers have been formed to deal with these issues. However, we assert that most of these problems can be solved with technical solutions and that these solutions mostly involve exploiting existing technologies such as cryptography, linked data and text mining. An important role is still to played however by these organizations in proposing and developing these solutions and promoting them within the relevant communities.

6. Acknowledgements

This research was supported by the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

7. Bibliographical References

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., et al. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227.

Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078.

Choukri, K., Arranz, V., Hamon, O., and Park, J. (2012). Practical and technical aspects for using the international standard language resource number. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 50–54.

Cohn, J. P. (2008). Citizen science: Can volunteers do real research? *BioScience*, 58(3):192–197.

Euzenat, J., Valtchev, P., et al. (2004). Similarity-based ontology alignment in OWL-lite. In *Proceedings of the 16th European Conference on Artificial Intelligence*, page 333.

GenomeCanada. (2005). Genome Canada data release and sharing policy. Technical report, GenomeCanada.

Goble, C. A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., et al. (2010). myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic acids research*, 38(2):677–682.

Groth, P., Gibson, A., and Velterop, J. (2010). The anatomy of a nanopublication. *Information Services and Use*, 30(1-2):51–56.

Klyne, G. and Carroll, J. J. (2006). Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation, World Wide Web Consortium.

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., and Zaveri, A. (2014). Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 747–758. ACM.

Maali, F., Erickson, J., and Archer, P. (2014). Data catalog vocabulary (DCAT). W3C recommendation, World Wide Web Consortium.

Maniatis, P., Roussopoulos, M., Giuli, T. J., Rosenthal, D. S., and Baker, M. (2005). The lockss peer-to-peer digital preservation system. *ACM Transactions on Computer Systems (TOCS)*, 23(1):2–50.

McCrae, J. P., Cimiano, P., Rodriguez-Doncel, V., Vila-Suero, D., Gracia, J., Matteis, L., Navigli, R., Abele, A., Vulcu, G., and Buitelaar, P. (2015). Reconciling Heterogeneous Descriptions of Language Resources. In *Proceedings of the 4th Workshop on Linked Data in Linguistics*, pages 39–48.

Monaghan, F., Bordea, G., Samp, K., and Buitelaar, P. (2010). Exploring your research: Sprinkling some Saffron on semantic web dog food. *Semantic Web Challenge at the International Semantic Web Conference*, 117:420–435.

Paskin, N. (2008). Digital object identifier (DOI) sys-

[6]Saffron: `http://saffron.insight-centre.org/`

tem. *Encyclopedia of library and information sciences*, 3:1586–1592.

Pinto, D., McCallum, A., Wei, X., and Croft, W. B. (2003). Table extraction using conditional random fields. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 235–242, New York, NY, USA. ACM.

Piwowar, H. A. and Chapman, W. W. (2010). Public sharing of research datasets: a pilot study of associations. *Journal of informetrics*, 4(2):148–156.

Prud'Hommeaux, E., Seaborne, A., et al. (2008). SPARQL query language for RDF. W3C recommendation, World Wide Web Consortium.

Troncy, R., Mannens, E., Pfeiffer, S., Deursen, D. V., Hausenblas, M., Jägenstedt, P., Jansen, J., Lafon, Y., Parker, C., and Steiner, T. (2012). Media Fragments URI 1.0 (basic). W3C Recommendation, World Wide Web Consortium.

Vossen, P., Bond, F., and McCrae, J. P. (2016). Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Global WordNet Conference 2016*.

Wellcome Trust. (1997). Wellcome trust statement on genome data release. Technical report, Wellcome Trust.

Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L., and Moore, A. J. (2010). Data archiving. *The American Naturalist*, 175(2):145–146.

Wilde, E. and Duerst, M. (2008). URI Fragment Identifiers for the text/plain Media Type. RFC 5147, RFC Editor.

Xu, W., Callison-Burch, C., and Dolan, W. B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in Twitter. *Proceedings of SemEval 2015*.