Data Commons and the Humanities

[1]Michael Haugh [2]Simon Musgrave

[1]University of Queensland, Brisbane, Michael.haugh@uq.edu.au
[2]Monash University, Melbourne, simon.musgrave@monash.edu

INTRODUCTION

A set of responses to the technological transformation of knowledge creation in the 21st century has developed around the idea of openness: Open Access, Open Science, Open Data. One part of these developments has been an increasing willingness to see knowledge and the infrastructure for its creations as a commons, that is, as a shared resource built, maintained and used by a community [1]. A crucial part of the infrastructure is a mechanism for making data widely accessible, a data commons. This concept has been adopted more widely in the sciences than in the humanities (see e.g. [2]). In this presentation, we suggest that this is the case because the requirements of a data commons in the humanities are different to those in the sciences. Although the scale of the data involved is smaller, the data itself is more complex and its uses may also be more complex, extending across multiple research communities. We discuss these points in relation to current plans for a Language Data Commons in Australia.

THE IDEA OF A DATA COMMONS

Borgman [3] identifies four benefits which come with sharing data: making replication possible, making the results of publicly-funded research accessible, making it possible to ask new questions of existing data, and advancing research and innovation in general. In all of these purposes, data is seen as a public good, and such resources fit well with the commons model of community ownership [4]. Establishing what is the relevant community and then setting up appropriate governance structures are nevertheless challenging questions.

HUMANITIES DATA

As Borgman pointed out [5], scholars in the humanities are not always comfortable with the term 'data'. But as more sources are made available as digital objects and as more scholars become comfortable in manipulating such objects with computational tools, it is hard to avoid thinking in terms of humanities data. An important difference between humanities data and data in other disciplines is that much less of the data is original or is created in the process of research. This aspect of humanities data only strengthens the arguments for it to be treated as a commons. Humanities data also differs from the data used in scientific disciplines in at least three ways (of course there are overlaps, but we believe the generalization is appropriate) which have consequences for how it is stored and accessed:

1. Volumes of data are typically smaller than those in 'big data' disciplines such as genomics. High resolution imaging and multimodal material are examples of humanities data which require storage space, but such resources do not form the largest part of humanities data.
2. Types of data are more varied. Humanities data can include resources such as text, images, audio and video, numerical data, and other possibilities. As will be discussed below, this variety occurs and is important even with in a single discipline. At the level of the humanities as a whole, it is an essential aspect of research activity.

3. Data is reuseable across disciplines. Reuse is a major aim of data sharing, but in many disciplines, such reuse is a meaningful opportunity only for researchers in the discipline. Making data accessible to interested parties outside the academic community can be a benefit (e.g. the use of Sloan Digital Sky Survey data by amateur astronomers [6]), but this is not commonplace. In contrast, humanities data is often relevant to scholars in a variety of disciplines as well as being of interest to non-scholars.

LANGUAGE DATA

The points made above in regard to humanities data generally are all relevant to the case of language data. Much research in linguistics and related fields today is data-driven. For example, considerable effort has been devoted to documenting the linguistic diversity of our species, much of which is threatened [7], and the results of such work typically include multimedia data. Spoken language is the basic data of linguistics and audio is therefore important; today video data is also very relevant for various research areas such as gesture, sign language and interaction research. In all cases, the amount of data collected and the amount available to be shared is constrained by the fact that such data is often of little value unless It has been enriched with annotations and providing such annotation is a time-consuming process. Text data is another type of language data, but such material is very economical to store. For example, the Corpus of Contemporary American English [8] is a collection of more than half a billion words of text data but the complete resource (with various types of annotation) takes less than 50GB of storage space. Data collected for linguistic research may be of value to researchers in other fields and vice versa. To give some obvious examples, a corpus of texts can be analysed by a linguist, but, depending on the content, the resource may also be of value to an historian, or a researcher in media studies or literary studies. On the other hand, material collected by an oral historian provides a linguist with data on language in use.

PROSPECTS FOR THE FUTURE

Most of the problems which are relevant to humanities data are already relevant to language data. The research community in Australia which is interested in language data is diverse but has some characteristics which suggest that efforts to establish a Language Data Commons would be fruitful. Although the community is diverse, it groups around three professional associations and shares many common aims. The developments in language documentation and the use of multimodal data in recent years mean that it is a community which is already comfortable working with digital data. The cost in time and money of data processing activities such as transcription and annotation means that sharing data is the best way to enlarge the scale of research with (some types of) language data. And much of the data involved can also be used by scholars in other disciplines and by members of the public. Efforts are already under way to create a resource of the type just sketched. The aim is to enable access to data which represents the linguistic situation in Australia in all its diversity, and this aim is best accomplished by aggregating assets rather than by a process of constrained data collection (as in conventional corpus building). Such access should be ethical – it should make access available as widely as possible while respecting the rights and sensitivities of those who have contributed to the data – and it should be equitable – the barriers in accessing the data should be minimised as far as is consistent with the ethical commitment. Therefore, our preferred model for the resource is a data commons.

REFERENCES

[1] C. Hess and E. Ostrom, Eds., *Understanding knowledge as a commons: from theory to practice*. Cambridge, Mass: MIT Press, 2007.

[2] R. L. Grossman, A. Heath, M. Murphy, M. Patterson, and W. Wells, "A Case for Data Commons: Toward Data Science as a Service," *Computing in Science & Engineering*, vol. 18, no. 5, pp. 10–20, Sep. 2016.

[3] C. L. Borgman, "The conundrum of sharing research data," *Acta Anaesthesiol Scand*, vol. 63, no. 6, pp. 1059–1078, Jun. 2012.

[4] J. Boyle, "Mertonianism Unbound?: imagining free, decentralized access to most cultural and scientific material," in *Understanding knowledge as a commons: from theory to practice*, C. Hess and E. Ostrom, Eds. Cambridge, Mass: MIT Press, 2007, pp. 123–144.

[5] C. L. Borgman, "The Digital Future is Now: A Call to Action for the Humanities," *Digital Humanities Quarterly*, vol. 3, p. 4, 2010.

[6] "SDSS." [Online]. Available: https://www.sdss.org/. [Accessed: 07-Jun-2019].

[7] J. Gippert, N. Himmelmann, and U. Mosel, Eds., *Essentials of Language Documentation*. Walter de Gruyter, 2006.

[8] M. Davies, "The Corpus of Contemporary American English: 520 million words, 1990-present," 2008. [Online]. Available: https://www.english-corpora.org/coca.