Digital humanities and language documentation

Nick Thieberger
School of Languages and Linguistics, University of Melbourne
thien@unimelb.edu.au

Language documentation is aligning linguistic methods with those more generally being adopted by Digital Humanists. Language documentation can be seen as delivering a new kind of postmodern linguistics by acknowledging the partiality of the data on which an analysis is based, and in explicitly inserting the linguist into the process of recording, annotating and preparing the corpus for scrutiny by others. Paradoxically, it is this kind of scientific method that also connects with other humanities disciplines under the rubric of Digital Humanities. There is a need for new ways to access and to analyse information with tools reflecting the methodological requirements of a particular discipline and a need for new ways to publish that information. These are the three foundational components of DH addressed in this paper. I discuss ways in which language documentation can benefit from practices and tools developed in the humanities and ask what is offered to language documentors by large digital infrastructure projects.

Keywords: Language Documentation, Digital Humanities
1. Introduction

Language documentation theory (LD) is aligning linguistic methods with aims more generally being adopted in the Digital Humanities (DH), especially in its focus on re-usability of primary research data. The model provided by descriptive linguistics in the past has mainly been concerned with writing linguistic analyses of narrowly focused phenomena and, occasionally, of lesser-known languages. More recently, a theory of language documentation has developed, partly in reaction to the abstraction of, for example, the minimalism of Chomsky and his followers, but also in response to the needs of people we work with in the field. In its focus on the creation of good records that can also be used by the people recorded, LD is addressing the criticism of earlier linguistic methods as being extractive, taking information but leaving little in the way of recordings or texts that could be of use for the source community.

LD can also be seen as delivering a new kind of linguistics by acknowledging the partiality of the data on which an analysis is based, and in explicitly inserting the linguist into the process of recording, annotating and preparing the corpus for scrutiny by others. It also emphasises the presentation of the context of an utterance in contrast with the earlier practice of basing a theoretical point on a decontextualised example sentence. All of this increases self-reflexivity and contingency in linguistic analysis, seemingly identifying it as a postmodern humanist enterprise, but, at the same time it is empiricist in building a set of data that can be used in a scientific method of data gathering, hypothesis development and testing and analysis and in which critical aspects are the replication of analyses, external review of collections and processes, and data recording and sharing.

DH has a focus on digitisation and creation of research infrastructure, and on the tools and methods required to create, describe, locate, annotate and curate the material that is the basis for much humanities research. There is some controversy within literary theory around the problem of quantification and the resulting
emphasis on scientific methods that is perceived as being part of DH (see e.g., Hall 2012 and Edwards 2012). While this discussion may have merit from the perspective of literary theory it does not apply to the kind of linguistic research discussed here, as linguistics has always been both a social science and a humanities discipline. In fact, as will be seen throughout this article, I suggest that the scientific basis for research claims that is provided by DH methods is a central benefit offered to LD.

The connection between DH and LD that is outlined in this paper is based on the shared need for tools and methods in the broader research community that can be applied in a researcher’s everyday work. There are common research practices in other humanities disciplines that could share solutions with LD, if we are able to abstract the functions we need from the content we work with, as sketched out in the points below.

- Primary annotation of primary documents (by researcher)
- Secondary annotation of primary documents (by other than the researcher)
- Use of repositories with discipline-based metadata
- Metadata-entry
- Data management methods
- Data capture
- Presentation of analysis with primary data

2. Digital Humanities

Manfred Thaller provides a useful definition of DH (or eHumanities): “eHumanities describes the concept of performing Humanities research in a distributed digital working environment, which supports equally well: (1) access to the information needed to tackle a research question, (2) the analysis of that information by tools reflecting the methodological requirements of the specific discipline and research problem and (3) the publication of the new information
gained by the analytical process.” (Thaller 2012:11). These three points will be addressed below.

There is a need for new ways to access and to analyse information with tools reflecting the particular methodological requirements of linguistics and LD and a need for new ways to publish that information. These are the three foundations of DH I will address below (§3.1). Thaller goes on to suggest that the provision of digital information is not enough, it must be accompanied by the kind of understanding that DH ‘defines and is defined by’ and is ‘[a]n understanding, that has to come out of the Humanities themselves, not the libraries – nor any other type of repository.’ (Thaller 2012:11) I take this ‘understanding’ to be the contextual information that makes the data interpretable, arising from its own disciplinary background.

2.1 The Companion to Digital Humanities

In order to come to grips with what DH may or may not have to offer LD we can look at the Companion to Digital Humanities (Schreibman, Siemens & Unsworth 2004), a collection of chapters relating DH to particular discipline areas. While it is a decade old now, it has chapters on several topics that are of potential relevance to LD. The section on lexicography focuses on paper-based monolingual dictionaries. It discusses the use of tools like TACT (DOS-based software for textual analysis) and concordancers, and notes the novelty of hyperlinks, showing its age. The linguistics chapter (titled ‘Linguistics Meets Exact Sciences’) deals with computational tools for NLP, machine translation, morphological parsers, POS tagging and applications in the minimalist program. It concludes that ‘computational linguistics has become an experimental science’.

Few of these techniques are of relevance to LD as they typically require training corpora that are orders of magnitude larger than a corpus normally produced from LD. Other potentially relevant chapters to LD, such as ‘Markup’, ‘Music’, ‘Text encoding’ and ‘Preparation of linguistic corpora’ do not deal with fieldwork data and so are not immediately useful but do provide theoretical guidelines for ways in which (in particular, textual) data should be prepared.
3. Language Documentation

LD has developed a following after being formulated by Himmelmann in his now well-known and often cited 1998 paper. For those outside of linguistics it may be useful to emphasise how challenging the ideas in LD are to the mainstream of linguistics because I think the challenge in linguistics mirrors similar challenges in other disciplines, and elicits similar negative reactions from established scholars.

LD is concerned to create records of the world’s small languages. Keep in mind that there are 7,400 languages in the world (Lewis, Simons & Fennig 2013) and that there are no records for many of them and very poor records for most of them. Since the 1950s many linguists have worked in a theoretical model which focusses on studying one’s own language by self-reflection. As a result little descriptive work on small languages was undertaken, and, when it was, it was often written within a theoretical framework that made it hard for outsiders to read, let alone for speakers of the language to use. Further, the creation of good primary records was never part of the descriptive approach, rather, the extraction of information by linguists was solely for the purpose of the first world academic.

Language descriptions are necessarily partial. They deal with an analysis based on what is fashionable in linguistic theory at the time. Each generation of linguists focuses on new features of language in the descriptions created. Recently we’ve heard about evidentiality (e.g., Aikhenvald 2004), but 20 years ago, nobody was talking about evidentiality. Mirativity (e.g., Hill 2012) suddenly came into vogue; we’re hearing a lot about referentiality now (e.g., Schnell 2012), and so on. We know that there are different fashions in analyses written at different times and that a grammar written before these times doesn’t say anything about what may be fashionable now. If, on the other hand, we had a grammar as the key to a rich accessible corpus, then that would be a much stronger set of material on which to build further analysis of the language, one that can also be used by others to substantiate our claims (see Thieberger 2009). Earlier linguistic fieldwork methods didn’t typically create good records. They were more concerned with expropriating information (in the sense of making records that were virtually impossible for the speakers or their families to access), analysis and grammar writing. If there were records (audio, notes, photographs), often they were left in the linguist’s office or home until they died and then somebody else had to deal
with them. And, partly as a result of this lack of a suitable methodology, grammatical descriptions didn’t make explicit the links between primary records and results of their analysis.

A recent example that is illustrative of the distinction between language description and documentation is that of the language Pirahã and claims made about it by Everett (2005), including that it has no numerals, very few kin terms, no colour terms, and no recursion in syntax. While Everett is a staunch opponent of Chomsky’s minimalism and of the notion of universals more generally, he uses a methodology in which the only language materials made available are the examples that he selects. And he is (or was when he began making the claims) one of the only linguists to know Pirahã, and the only one with access to Pirahã recordings. If he had established a collection in a repository that provided the necessary framework to locate his records and the licencing to allow its use, then others could test his claims against that documentation.

Much has been made of the difference that LD represents as a methodology in linguistic research, but, in some ways LD takes us back to the early twentieth century anthropological linguistics of Boas and Sapir, with the emphasis on texts and dictionaries in addition to grammatical analysis. A major difference is that new tools allow linguists to qualitatively change the way we work with primary data and to increase the number and size of collections of recordings. As long ago as the 1930s Malinowski noted the possible benefits being offered by new technologies:

‘If I could, by a good phonographic record, counterfeit the living voice of Tokulubakiki: [...] I should certainly be better able to translate the text in the sense of imparting to it its full cultural flavour and significance. Again, if by cinematographic picture I could reproduce the facial expression, the bodily attitude, the significant gestures, this would add another contextual dimension [...] There is no reason whatever why, in the future, an exact and physiological study of speech should not use the apparatus of sound films for reproducing fully contextualised utterances.’
(Malinowski 1935: 26)

While there may have been no reason, it didn’t happen—since 1935 we have not had within fieldwork-based linguistics a method for using audio recordings as the basis for an analysis. In fact, there have been a few rare attempts to present
primary recordings with texts, and to explore the use of hypertext for opening links between texts, grammar, lexicon and media. In 1970 von Brandenstein, a linguist working in Western Australia, published a 45 rpm record with his collection of texts in Ngarluma and Yindjibarndi (Brandenstein 1970). In the late 1980s Valentine (1992) produced a set of HyperCard stacks that presented audio and a transcript of a text in Ojibwe\(^2\), together with morphemic glosses that, when clicked, were links to a dictionary, and part of speech information that linked to a grammar sketch. This was an elegant model that has not since been taken up.

Why has it taken so long for such ideas to become actuality? In part it has to be acknowledged that we are all tied in to methods that we are brought up with and are likely to be slow to change. The tools to do this work have been difficult to use. And there has been a lack of leadership among some of our more senior colleagues. Furthermore, we can see affordances more than we can see solutions, that is, the possibilities offered by new technologies remain largely unrealized because tools and methods are still in development. The task in both DH and LD is to model what can be done and to show that benefits flow from adopting new methods. Training in the use of tools is being carried out, in particular by funding programs like DoBeS and ELDP, but also by intensive summer schools like CoLang\(^3\). There is a consensus in LD in general that certain workflows exist (cf. Thieberger & Berez 2012:97) and that a limited set of tools provide the best way to process and analyse field recordings.

In modern research the use of computers is central, but without proper understanding of the nature of the data created, much digital data will end up being lost or created in formats that are not available even to the creator within a short period of time. That is, digital creation of humanities research material is not, in itself, sufficient to constitute DH research. In LD such loss is particularly problematic. Dictionaries created in MS Word are a good example. The DH approach would promote the use of lexicographic software to build a lexical database from which a dictionary, a reversal and topical lists can all be derived.

On the other hand dictionaries written exclusively with their appearance as a book


\(^3\) [http://tinyurl.com/colang2014](http://tinyurl.com/colang2014)
in mind may encode structural elements with formatting changes (bold headwords, italic part of speech and so on) and, what’s worse, there is then a transition to a page-layout program where further corrections are made and not reflected back into the original documents. Once the book has been produced, the source files have typically not been kept. As a result, subsequent editions become laborious and time-consuming. Reversals, or finderlists, can, in principle, be constructed from the same underlying lexical database as the dictionary, but this can’t be done if there is no lexical database. Hence, a dictionary is produced and printed as a fine book, but the reversal is extra labour that remains to be done, as are other derived forms of the lexical data, such as topical lists, web-based dictionaries or phone apps.

This understanding of the nature of text, and the use of structured text to allow multiple outputs from a single source, is also central to DH practice and is a point at which the practice of LD could inform the general humanities community.

Having outlined a working definition of DH and LD, the following sections discuss how Thaller’s three defining features of DH relate to LD.

### 3.1 Access

The innovation that LD brings to accessing language records is in its development of digital archives. These archives have immediate relationships with researchers, and in particular they can provide guidance and training in the creation of records that can be archived relatively easily. As longevity of the records produced by fieldworkers is central to LD, archives are now integral to the LD workflow, and are no longer relegated to sometime after the researcher’s retirement or death.

Language archives have developed an international community of practice over the past decade and typically provide their catalogs for aggregation by the Open Language Archives Community (OLAC). This allows OLAC to list what is available for each of the world’s languages via a url, http://www.language-archives.org/language/XXX, in which the last three characters are the ISO639-3 code for a language. As each archive updates its catalog so too does this page-per-language get updated. Linguists have been at the forefront of developing
repositories for primary research outputs, something that will certainly be part of DH as it develops but is not yet established for most humanities disciplines.

3.2 Analysis and Tools

Software tools for LD are the fundamental determinants of the data structures used by its practitioners. This can be seen in the software produced by the Summer Institute of Linguistics\(^4\) (like Toolbox and Fieldworks), and by MPI Nijmegen\(^5\) (Elan\(^6\) and Arbil\(^7\)), which is designed for linguists and is based on principles of open-formats and structured data. In Australia, HCS vLab (discussed below) has created a set of tools and processes on top of collections of primary data, allowing concordancing, lemmatising, tokenization and other ways of visualising and working with large sets of heterogeneous primary records.

3.3 Publication

The third part of our definition of DH is the publication of new information—especially using new means of delivery—an area where LD is falling behind a little. We don’t currently have publication methods for language corpora beyond sending around source files in non-standard formats. For example, a corpus of texts in Toolbox has no guarantee of being useable by anyone other than the original creator, unless they have documented the fieldnames they are using and provide the associated files required by Toolbox to make sense of the structure of the texts. In a step toward making interlinear corpora accessible I developed EOPAS\(^8\), an online corpus presentation system which takes Toolbox or ELAN annotated texts and media and presents them online. Texts in EOPAS are citable at the level of morphemes or utterance units and allow users to access media

\(^4\) http://www.sil.org
\(^5\) http://www.mpi.nl
\(^6\) tla.mpi.nl/tools/tla-tools/elan/
\(^7\) tla.mpi.nl/tools/tla-tools/arbil/
\(^8\) http://www.eopas.org
time-aligned to the text. EOPAS is open source⁹ and an ideal future development would be to allow online annotation of media.

Often linguistic research involves looking through grammars to find what they say about a particular feature. If the author did not mention a feature in the language then subsequent researchers want to search for information that isn’t in the grammar. How can that information be found? If there were a collection of annotated texts we could search through them, but how do we find the texts? Language archives are emerging as publishers of primary material (see the discussion about PARADISEC below) but that can only work if data structures are being created suitably to allow the materials (recordings, annotations, images, analysed texts, lexicons) to flow into archives and perhaps on to the web and other publication forms. Making a corpus of data available through an archive is a new, exciting, and important mode of publication, but it is also a way of making data accessible to wider research communities (and other users). The ability to cite primary material will become increasingly relevant as more repositories are created and their use is encouraged or mandated by scholarly publishers and funding agencies.

A basic problem for the creation of research material in LD is the inclusion of contextual information about the process of data creation. Typically in the course of fieldwork, but possibly also in archival research on historical sources, linguists understand the context in which the data is created, and create the analytical apparatus around it, so are then able to describe it and they have to be involved in describing it and preparing it for curation. However, at a certain point, such curated material enters into a repository and becomes useful for others, and for potentially very different uses than the creators had as their goals. Thus, while the DH/LD analytical apparatus is required for a particular interpretation of the records, there should, as a matter of principle, be a clear separation of the primary records from that apparatus so that the records can take on their own life independent of the original creator. Ontologies, or standardised lists of terms that can be mapped to from each particular instance of a collection will, in time, allow the products of each of the various research communities to interoperate.

⁹ https://github.com/eopas/
4. The Australian context for LD and DH

In Australia there is a newly established Australasian Association for the Digital Humanities\textsuperscript{10}, there have recently been positions advertised at the ANU and the University of Tasmania identified as being in DH, and the first chair in DH has been appointed at the University of Western Sydney. There is also a long tradition of descriptive linguistics with an active group of linguists interested in DH.

There are different models of how DH units can operate, but two that have been successful are at Kings College London and the Cologne Centre for eHumanities which both operate as partners for scholars in the humanities, providing advice and technical support to ensure that research is done using appropriate methods. At the University of Melbourne the eScholarship Research Centre (eSRC) operates as a unit of the library and collaborates in projects with academics. I have worked with the eSRC to prepare a set of some 14,000 manuscript page images from deceased estates of three linguists (see for example the work on Arthur Capell’s manuscripts here: http://paradisec.org.au/fieldnotes/AC2.htm).

5. My practice

In my own work I have been very interested to build methods for preparing language records in ways that can be reused, and in the research toward my grammar of the language of South Efate (Vanuatu) in the late 1990s I wanted to ensure that all examples were cited to a media and transcript source. I had to develop a method to do this, as it was not part of any existing linguistic fieldwork practice. I did not understand how generalizations could be made on the basis of field recordings with no method for accessing those recordings and for citing points within them. As there was not a suitable tool I wrote one in HyperTalk, the only programming language I knew (Thieberger 2004). \textit{Audiamus}\textsuperscript{11} is a tool I still use to access my media data via a transcript, although more recently this would be done using ELAN. \textit{Audiamus} holds transcripts of some 23 hours of media, from

\textsuperscript{10} http://aa-dh.org/

\textsuperscript{11} http://languages-linguistics.unimelb.edu.au/thieberger/audiamus.htm
which I have selected sections for further annotation with morphemic glosses. I had no funding to pay anyone to write this software for me, and I doubt I could have given anything but the barest outline of what I wanted it to do at the outset.

As it developed I kept adding features: a concordance; the ability to create playlists; a citation button that clips a piece of text and its media name and offsets. In other programming/development efforts since then I have found this to be a useful approach, building rapid prototypes, and using wireframes to see how a tool will work before it is fully developed. With Audiamus I was able to build a media corpus and to cite points within it in my PhD dissertation, and then harvest all these references once the thesis was finished to make a playable standalone version of Audiamus that was presented on a DVD, presented also in the published book version. While some colleagues regard this as a gimmick, they are decreasing in number in the decade since my PhD was completed. The basic principle is that data should be provided to allow replication of analyses and that it has to be digital in order to do that. The research methods made possible by the use of Audiamus, to recall Thaller’s three definitions of DH mentioned earlier, have provided access to recordings, both via a grammatical analysis and as a corpus and have allowed publication of that corpus for review and validation by other scholars. So, from a strictly ergonomic and even economic point of view, DH offers solid research methods on which subsequent research can build.

6. Citation of primary data: Language archives and the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

As an academic community we have created methods for citation of our outputs—journal articles, books and so on. After several hundred years of practice there is a well accepted convention for building on existing research and for evaluating impact factors, with consequences for promotion and career advancement. We have a little more trouble with citation and evaluation of web-based documents. Eprints repositories deal with online access to what has until now been normal research output, that is, papers, articles and books, but they do not, in general, deal with primary research data. We (and here I refer to linguists) still have a long way to go with citation of primary records. In order to provide ongoing access to
primary records they need to be held in archives with persistent locations and names.

PARADISEC (Thieberger & Barwick 2012) is an archive I helped establish a decade ago to address precisely this issue of the lack of attention paid by linguists to their own primary data. Because the old method of linguistic fieldwork was almost exclusively focused on writing grammatical descriptions any primary recordings made by a linguist were regarded as not having any intrinsic value. This can be seen by the way in which they were treated – left in filing cabinets, in boxes in garages, in deceased estates, with no catalogs of their contents. The community of linguists through to the end of the twentieth century did not, collectively, make provision for these collections to be curated12.

PARADISEC’s collection is structured sufficiently to allow records to be located via standardised metadata descriptions. As much of the collection is made up of legacy recordings the metadata is limited and will eventually benefit from the exposure of the primary records for eventual online annotation. Such exposure is provided via metadata feeds from our catalog13 which are harvested by a number of different services, including, most recently, two Australian Humanities infrastructure projects, Humanities Networked Infrastructure (HuNI14) and HCS vLab15.

These large projects help to show what is possible by building new ways of relating research outputs and analysing them, thus enticing researchers to create their collections in ways that will be useable by the new infrastructure.

---

12 An exception to this general observation is that the Australian Institute of Aboriginal and Torres Strait Islander Studies provides a repository for recordings related to Australian languages. Grantees of AIATSIS are obliged to deposit their recordings, but, even then, it is not the case that all grantees have deposited their recordings.

13 http://catalog.paradisec.org.au/apidoc

14 http://huni.net.au/

7. Conclusion

An old man who I worked with in Erakor village (Efate, Vanuatu) told a story which had as its moral that “a small stone can break a big canoe”\(^\text{16}\), suggesting that even the largest enterprise can fail due to small stumbling blocks. It is clear that there are a number of small stones breaking the big canoe of digital research methods at the moment. Many of these small stones strewn over the research landscape come from the lack of understanding of the issues and manifest as very simple problems, like poor file naming practices, lack of metadata, or poor choices of software. Larger problems include the lack of generic tools for similar data management and analysis processes, the lack of suitable repositories for primary research outputs and the lack of recognition of primary data collections as having any research value. All of these problems are shared by linguists and the broader DH community.

Humanities research data—such as historical records, information about local languages, or indigenous stories—is typically of interest to the general public in ways that most science data is not. Preparing this material for re-use should be a priority for cultural agencies and funding bodies, but the task of digitizing it and preparing it for re-entry into the online research environment is typically not seen as worthy of funding, while at the same time vast amounts of funds are made available for national data services in European projects like DARIAH\(^\text{17}\) and Clarin\(^\text{18}\), or Bamboo\(^\text{19}\) in the USA, and HuNI and HCS vLab in Australia. These large scale projects risk having little content if the more important data preservation task of digitising analog media is not also funded.

Once the data is online via a repository then it can be made available for crowdsourcing annotations, for example, putting linguistic manuscripts online using this framework, and then allowing other people to do annotations to enrich

\(^{16}\text{This story can be read and heard here: http://www.eopas.org/transcripts/80#t=1770.86,1782.14}\)

\(^{17}\text{http://www.dariah.eu/}\)

\(^{18}\text{http://www.clarin.eu/}\)

\(^{19}\text{http://www.projectbamboo.org/}\)
the collection. In this way, LD could build on more generic DH methods for doing this (see for example Scripto\(^20\)).

Another of the small stones breaking the large canoe of LD is the lack of metadata provided by researchers for their research collections, in part because there are no simple metadata entry tools. There are several such tools available now, including Arbi\(^21\) (mainly used by DoBeS grantees), SayMore\(^22\) (Windows only) and ExSite\(^9\)\(^23\) (a bare-bones metadata entry system) but none of them has had much take-up outside of their originating communities.

Beyond metadata entry, we need: cataloging software for archives; simple methods for citation of points within primary media; standard formats for basic types of texts like interlinear glossed text, which is such a common format for linguists to use, but for which we don’t have a standard format (see Schmidt 2003, and Bow, Hughes & Bird 2003). And, critically, we need more digital repositories for research outputs.

In summary, Language Documentation is a natural part of the Digital Humanities in its emphasis on the cumulative nature of research. Humanists need local centres for advice and support, and language documenters are in a good position to provide advice. Linguists, who routinely work with structured text, have an understanding of the ways in which it can be converted into various output formats, and made to work in visualisation tools. There are advantages for language documentation in participating in large digital humanities projects but that is predicated on LD adopting the right methods and training new practitioners to understand why all of this is important.

References


\(^{20}\) http://scripto.org/
\(^{21}\) http://tla.mpi.nl/tools/tla-tools/arbil/
\(^{22}\) http://saymore.palaso.org/
\(^{23}\) http://www.intersect.org.au/exsite9
Thieberger


Brandenstein CG von 1970 Narratives from the North-West of Western Australia in the Ngarluma and jindijparndi languages (Volumes 1–3 + audio-disc) Canberra: Australian Institute of Aboriginal Studies.


Schnell S 2012 ‘Data from language documentations in research on referential hierarchies’ in F Seifart G Haig NP Himmelmann D Jung A Margetts & P Trilsbeek (eds) Potentials of language documentation: Methods, analyses, and utilization Language Documentation & Conservation Special Publication No. 3. pp. 64–72


Author/s: Thieberger, N

Title: Digital humanities and language documentation

Date: 2014

Publication Status: Published

Persistent Link: http://hdl.handle.net/11343/40961

File Description: Digital humanities and language documentation