White Paper Report

Report ID: 105657

Application Number: HJ5009212

Project Director: Colin Allen (colallen@indiana.edu)

Institution: Indiana University, Bloomington

Reporting Period: 4/1/2012-1/31/2014

Report Due: 4/30/2014

Date Submitted: 4/29/2014

From Big Data to Argument Analysis:

A Selective Study of Argument in the Philosophy of Animal Psychology from the Volumes of the Hathi Trust Collection

authors

Simon McAlister, Colin Allen, Andrew Ravenscroft,
Chris Reed, David Bourget, John Lawrence,
Katy Börner & Robert Light

(with thanks to the full team at digging by debating)

Digging by Debating: Linking massive datasets to specific arguments

Final Report and Research White Paper

April 2014

Table of Contents

Abstract

The Digging by Debating (DbyD) project aims to identify, extract, model, map and visualise philosophical arguments in very large text repositories such as the Hathi Trust. The project has: 1) developed a method for visualizing points of contact between philosophy and the sciences; 2) used topic modeling to identify the volumes, and pages within those volumes, which are 'rich' in a chosen topic; 3) used a semi-formal discourse analysis technique to manually identify key arguments in the selected pages; 4) used the OVA argument mapping tool to represent and map the key identified arguments and provide a framework for comparative analysis; 5) devised and used a novel analysis framework applied to the mapped arguments covering role, content and source of propositions, and the importance, context and meaning of arguments; 6) created a prototype tool for identifying propositions, using naive Bayes classifiers, and for identifying argument structure in chosen texts, using propositional similarity; 7) created tools to apply topic modeling to tasks of rating similarity of papers in the PhilPapers repository. The methods from 1 to 5 above, have enabled us to locate and extract the key arguments from each text. It is significant that, in applying the methods, a non-expert with limited or no domain knowledge of philosophy has both identified the volumes of interest from a key 'Big Data Set' (Hathi Trust) AND identified key arguments within these texts. This provided several key insights about the nature and form of arguments in historical texts, and is a proof-of-concept design for a tool that will be usable by scholars. They have further created a dataset with which to train and test prototype tools for both proposition and argument extraction. Though at an early stage, these preliminary results are promising given the complexity of the task.

Specifically, we have prototyped a set of tools and methods that allow scholars to move between macro-scale, global views of the distributions of philosophical themes in such repositories, and micro-scale analyses of the arguments appearing on specific pages in texts belonging to the repository. Our approach spans bibliographic analysis, science mapping, and LDA topic modeling conducted at Indiana University and machine-assisted argument markup into Argument Interchange Format (AIF) using the OVA (Online Visualization of Argument) tool from the University of Dundee, where the latter has been used to analyse and represent arguments by the team based at the University of East London. This work has been articulated as a proof of concept tool – linked to the repository PhilPapers – designed by members linked to the University of London. This project is showing for the first time how big data text processing techniques can be combined with deep structural analysis to provide researchers and students with navigation and interaction tools for engaging with the large and rich resources provided by datasets such as the Hathi Trust and PhilPapers. Ultimately our efforts show how the computational humanities can bridge the gulf between the "big data" perspective of first-generation digital humanities and the close readings of text that are the "bread and butter" of more traditional scholarship in the humanities.

1. Introduction and Rationale

The Digging by Debating project aims to extract, model, map and visualise argument from a Big Data repository, such as the Hathi Trust Digital Library[1] (see Digging by Debating Proposal). Digging by Debating is part of the Digging into Data (Round 2) exploration of Big Data exploitation research. Preliminary studies conducted by team members, using various philosophical texts from Hathi Trust Digital Library, PhilPapers and other philosophy sources, demonstrated the challenge of developing a suitable mark-up methodology and then mapping arguments within an existing computational modeling scheme (the Argument Interchange Format[2]).

This project tackles the problem through:

1. linking the topic modeling work performed at Indiana University to a process for topic and sample selection of text destined for argument analysis from a Big Data set (the Hathi Trust);
2. applying Design based research methods (at ICPuP, UEL), including discourse analysis, argumentation mapping and other empirical techniques to problematise the design space, understand the nature of 'argument as data' and provide principles for tool design;
3. performing systematic human mark-up and analysis across a defined sample of texts using OVA Plus[3] as a mark-up tool, to provide representations and insights that could link the design based research work (at UEL) to the (Dundee) work into the automatic identification and extraction of arguments;
4. using statistical techniques to identify textual expressions of individual propositions, and then structural analysis techniques building on the University of Dundee's Semantic Web standards for argument representation to take initial steps in automatically identifying argument structure.

It was anticipated that the outcomes of part 2) should allow us to decide whether relevant arguments can be suitably mapped using OVA, and therefore support and inform the development of automated analysis algorithms to do the same; or, understanding and mapping arguments in our texts is particularly complicated and requiring complex human interpretation, and therefore tool development should also focus on interfaces to better understand, identify and represent arguments; or, whether tool development should consider both options.

Another way of looking at this 'argument study', from a broad cognitive science perspective, is that we are using formal modeling techniques and an empirical strand to better understand the nature of argument in the texts we're interested in – problematising the design space. Using this empirical strand to support and inform tool development should allow us to avoid

[1] The Hathi Trust http://www.hathitrust.org/about

[2] AIF see http://www.arg.dundee.ac.uk/aif

[3] See reference to OVA http://ova.computing.dundee.ac.uk

the common pitfall in this type of project, that of trying to solve a problem not properly understood. This also naturally brings in deeper issues about Digging into Data, as we are taking a closer look at the processes of meaning making within, and around, our ‘raw documents’ from an argument or argumentation perspective. Big Data projects can suffer from opaque conceptualisations of the data under examination and arguably too much focus on what can be done with it. We hold that, in order to usefully dig into data, we need to understand it to begin with.

Another related aim of this study is to use human understanding, as far as possible without preconception, to determine the nature of argument in indicative volumes dealing with this Animal Psychology topic. It is hoped that such an understanding will help to bridge topic modeling and argument selection with argument mapping and automated argument analysis, so findings will be usefully fed back into, and improve, a machine analytical study.

1.1 The Data of Interest: Animal Psychology 1880-1910

The era of the late 19th century and early 20th century was a period of significant development for psychology. Experimental methods were on the rise, and psychologists, who had often been housed in the same university department as the philosophers, were professionalising, forming their own associations and journals, and their own departments. Philosophy could be seen as in retreat from the advances of science, in particular the evidence-based, experiment-based scientific arguments favored by psychologists. At the same time, psychologists were wondering which of their received concepts and theories should be jettisoned, and which could form the basis of further progress.

Such questions were particularly acute in animal comparative psychology. On the one hand, Darwin’s theory of evolution exerted a strong pull towards the idea of mental continuity between humans and animals. On the other hand, many Darwinians were seemingly content with anecdotal evidence of animal intelligence to make their case on analogical grounds to human behaviour, leading experimentally inclined psychologists to reject such anecdotes and analogies as “anthropomorphic”. Even as the disciplines of psychology and philosophy were formally disassociating themselves, philosophical arguments about the “proper” way to study animal psychology were becoming even more prominent among the psychologists themselves.

We chose to focus on this era and these debates because (i) it matched the expertise of one of the P.I.s (Allen) and the academic background of the other P.I. (Ravenscroft); (ii) it offered a rich vein of philosophical arguments in the context of science; and, (ii) due to copyright limitations, we had full text access from the Hathi Trust Research Center (HTRC) only to texts prior to 1928, which nevertheless adequately covered the most important period from Darwin to the birth of behaviorism in psychology. *Our goal was not just to provide high-level summaries of the discussions going on in this era, but to see whether we could provide a method to help scholars identify specific passages containing highly relevant argumentation on the topic of animal mind.*

The particular difficulties of understanding arguments from this historical era are: a) not all the content is congruent with the style of scientific thought and writing that we have come to expect in the modern era (e.g., the heavier reliance on anecdotal evidence in earlier times); b) the language used is indirect, and verbose compared with its modern-day equivalent (e.g. there may be long digressions), and c) what passes for acceptable argument may well have been different in that era (e.g. rhetorical argument was much more common). So, when we consider that the language is difficult and verbose, the standards of 'acceptable' evidence low, and what is regarded as argument is different from our own understanding, we can foresee that at times the machine analytical task of determining passages containing argumentation will be working in an unfamiliar landscape with few landmarks – so scoping the difficulty of this was precisely what an empirical argumentation study could address. However, these difficulties will not always pertain, as there are volumes that are written with a scientific basis, even if the wording seems indirect and ponderous for the modern day. It is worth noting that the most successful attempts at automated argument modeling to date (Moens, et al. 2007; Palau & Moens 2011; Merity et al. 2009) have used scientific articles in a modern, formulaic structure (i.e., with ‘Introduction’, ‘Results’, ‘Conclusions’ etc., explicitly identified). The older texts in the collection that we built for this project do not have this structure. The task of understanding, identifying and mapping arguments in more ‘free running’ social science or philosophical (and historical) texts could be considered an 'order of magnitude' more challenging, given the great variety of styles, content and approaches.

A key innovation of our approach is that we adopt a multilevel approach to content analysis, which we briefly describe here before giving more detail in later sections of this report. As a first pass in locating the materials we wanted, we extracted a list of 1315 books in the Hathi Trust Digital Library that matched some known keywords. We then modeled these 1315 volumes using Latent Dirichlet Allocation (LDA) topic modeling, treating each book as a document. The resulting LDA topic model was scanned by a person who selected thresholds on the topics to extract 86 volumes automatically from the original 1315, as those most closely related to our interest in anthropomorphism and animal minds. These 86 were remodeled using LDA topic modeling with every page in these books treated as a document. A further step of topic-model assisted selection yielded six books of central interest for our argument analysis. As a final modeling step, we trained LDA topic models by taking each of these six books individually as a corpus, and every sentence in a book as a ‘document’.

The six volumes selected by the methods described in section 2.2 each focus (to some extent) on our chosen topic of Animal Psychology. The indicative volumes are (1880-1910):

1. THE ANIMAL MIND: A Text-Book of Comparative Psychology by Margaret Floy Washburn, 1908
2. COMPARATIVE STUDIES: Psychology of Ants and of Higher Animals by Eric Wasmann, 1905
3. THE PRINCIPLES OF HEREDITY by G. Archdall Reid, 1906
4. GENERAL BIOLOGY by James G. Needham, 1910
5. THE NATURE & DEVELOPMENT OF ANIMAL INTELLIGENCE by Wesley Mills, 1898
6. PROGRESS OF SCIENCE IN THE CENTURY by J. Arthur Thomson, 1908

2. Methodology: Linking Large Scale Visualisation, Topic Modeling and Argument Analysis

Semantic modeling of the entire Hathi Trust collection of more than 2.5 million books is currently not computationally feasible. By choosing instead to focus on philosophical disputes about animal psychology in the late 19th and early 20th C., playing to expertise within the Digging by Debating group, we were able to focus on a period of intense activity during the separation of psychology as a scientific discipline from its previous home in philosophy. In the subsections below, we describe in detail the methods we developed to move from a high level overview of 'too-large a collection to read', to finding relevant arguments of specific interest to our declared research topic of 'animal mind'.

2.1 Large-Scale Visualisation of Science-Philosophy Interactions

This part of the project aimed to empirically measure and visualize the cross-pollination of sciences and philosophy through paper citation data. Using the 1,400 plus articles from the Stanford Encyclopedia of Philosophy (SEP) as a proxy for the philosophical literature and a subset of 1,315 books from the HathiTrust scanned books collection (see data details above), we plot SEP citations (i.e. materials cited by SEP articles) and HathiTrust books onto the UCSD Map of Science to highlight areas of science which overlap with philosophical discussion. Do philosophers pay more attention to biology or physics? Geology or anthropology? Scientometric and text mining methods can suggest hypotheses in the early stages of an investigation. Subsequently, we describe the UCSD Map of Science and Classification system, the alignments used to overlap SEP articles and HathiTrust books onto the UCSD map, interpret results, and discuss next steps.

2.1.1 Method

The USCD map of science and classification system was previously computed by Börner et al. (2012) using 10-year article level data from Thomson Reuters' Web of Science and Elsevier's Scopus. It organizes major scientific journals into 554 subdisciplines (e.g., Contemporary Philosophy, Zoology, Earthquake Engineering) that are further aggregated into 13 core disciplines (e.g., Biology, Earth Sciences, Humanities). These sub-disciplines were laid out in a two-dimensional space to facilitate the overlay of other information. Journals and their papers that are considered more similar to each other are in closer proximity.

Each of the 554 subdisciplines has a set of journals and keywords associated. A new dataset, e.g., a set of publications by a researcher or institution, is overlaid by matching journal names to the journal names associated with the 554 subdisciplines. Fractional counting is used for highly interdisciplinary journals. Non-journal data (e.g., books, patents, funding, job advertisements) can be matched based on keywords, or by creating a crosswalk.

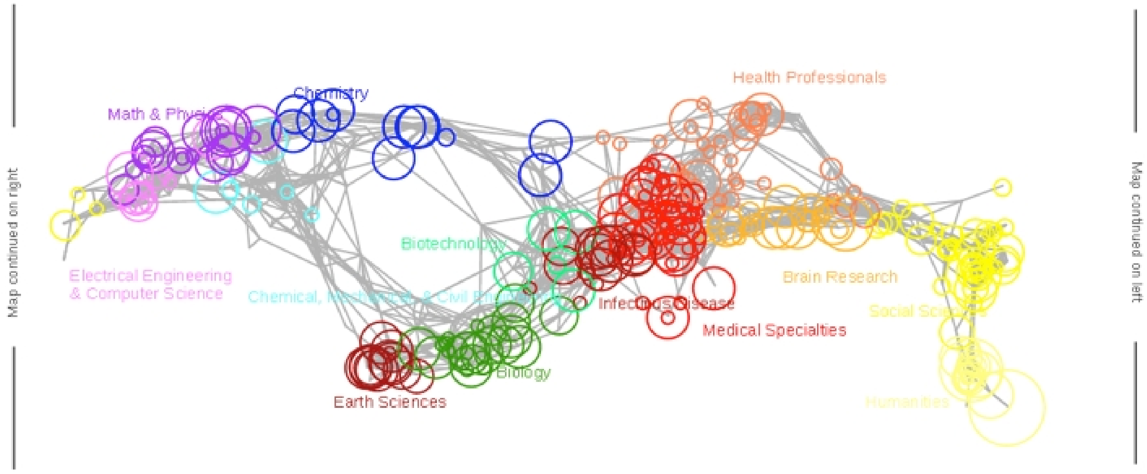
The SEP citations were mapped as follows onto the UCSD map. For each subdiscipline represented in the UCSD map, one can compute the number of SEP citations to the journals that make up that subdiscipline. Thus we can measure the influence of various areas of science in the SEP. By overlaying a representation of the number of citation hits from the SEP bibliographies into the various regions of the UCSD map, it is possible to visualize which areas of science have received relatively more or relatively less attention from philosophers (see detailed documentation in Murdock et al. 2013).

In order to overlay HathiTrust books, a crosswalk was created from the Library of Congress Call Numbers (LCCN) to the UCSD Map. To create the UCSD-LoC map, 2,010 journals with LCCN data provided by Andrew Asher of Indiana University's Wells Library were mapped by title to the UCSD Map of Science, with the number of titles that share a LCCN and a UCSD subdiscipline giving weight to each linkage. The 776 volumes from the 1,315 volume corpus of Hathi Trust materials that included LCCNs were mapped from Library of Congress to UCSD as follows. Those volumes that shared an LCCN with some number of mapped journals were directly mapped based on the location of those journals. However, many journals are assigned LCCNs numbers specifically designated for periodicals, so a perfect match at the topic level is rare. In cases where a direct match was not possible, the algorithm iterated up the hierarchy, taking the most narrow subsection that included the unmatched numbers as well as some number of mapped journals. Once a match was found, the location of the volume on the map was determined to be the center of all the LoC subsections that contributed to that mapping (one in the case of an exact match). For example, one of the 1,315 books with the title Evolution has a call number of B818 (Evolution, Holism), but as no journals in the USCD map of science share that call number, it is instead mapped as B808-849 (Special topics and schools of philosophy). The location of each subsection is the average of the locations of the subdisciplines to which it maps. Carrying the previous example further, four journals map in the B808-849 range. Three of these map to Philosophy/Psychology, while the other maps to Contemporary Philosophy. Since we do not weight by the number of journals mapping to a specific subdiscipline, the location for B808-849 and any journals that map to that location is the midpoint between the points for Philosophy/Psychology and Contemporary Philosophy. The resulting map is shown in Fig. 2.

2.1.2 Challenges

The effort to map the HathiTrust corpus to the Map of Science faced a number of obstacles, many unexpected. To begin with, finding a full description of the Library of Congress Classification System proved to be very challenging. While brief overall descriptions are easy to locate, and some individual sections are written up, a complete, machine readable file describing the entire classification structure proved elusive. One was finally located, but even this was incomplete, inconsistent and had errors. Formatting problems, uneven data descriptions and even a circular reference plagued the set, requiring a great deal of manual correction before this file could be used to power an automated mapping tool. Even once the resources required to perform the mapping were in place, only slightly more than 60% of the relevant volumes had the metadata required for mapping included with the hathiTrust record, and for those texts without a Library of Congress Call Number, the devised crosswalk could not be used. Therefore, outside sources had to be consulted to gather the needed metadata and complete the map.

2.1.3 Results and interpretation



©2008 The Regents of the University of California and SciTech Strategies.
Map updated by SciTech Strategies, OST, and CNS in 2011.

Figure 1. *Philosophical content mapped onto the UCSD Map of Science. The size of each circle corresponds to the number of SEP editorial areas citing material from the UCSD Map of Science subdiscipline (minimum: 0, maximum: 43). The color of each circle corresponds to one of 13 disciplines identified by the map of science. (Source: Murdock, Light, Allen, and Börner, JCDL 2013).*

The resulting maps are shown in Figures 1 & 2. Figure 1 indicates that SEP articles contain citations to journals that are very broadly distributed across the sciences, although the engineering fields (aquamarine region) appear to receive relatively less attention. Figure 2 confirms that the initial keyword-based selection of 1,315 books is more targeted towards fields below the “equator” of the map, with particular concentrations in the life sciences and humanities, as was to be hoped. The map provides additional visual confirmation that the further selections via topic modeling to a subset of 86 and then 6 of the original collection of 1,315 managed to target books in appropriate areas of interest.

Currently, the only validation of the alignments is by manual inspection. This is obviously an area for future work.

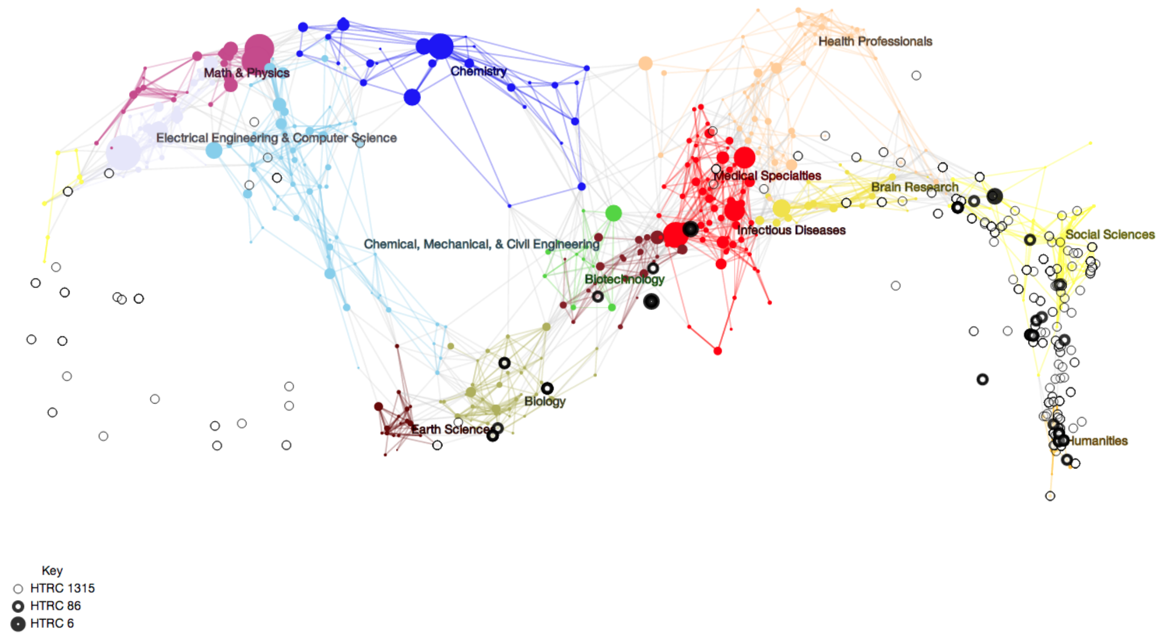


Figure 2. The HTRC 1,315 Corpus. *This data projection for the UCSD Map of Science shows each volume in the InPhO's HTRC 1,315 corpus on* ***anthropomorphism****. In the online interactive version hovering over the black dot displays the title of the volume, and clicking the dot will link to the volume in the HTRC (with some exceptions). The scroll wheel can be used to zoom in and out. There are currently 776 volumes on the map due to incomplete HTRC metadata. Books belonging to selected subsets are indicated with darker circles as shown in the key, bottom left.*

2.1.4 An interactive interface

A prototype interactive interface has been set up[4] that shows the UCSD map of science with proportional symbol overlays of the 1,315 HathiTrust books. Nodes can be selected, showing which volumes are mapped and providing the title and links to various external sources of metadata. The goal is that this type of grand overview can be used to select specific areas that lead to articles or books, and provide the tools to perform more intensive analysis of those texts. For example, an ambition is to use the visual overview provided by the Science-Philosophy map as an 'interface' for identifying and selecting topics of interest that can then be searched for across collections to identify particular texts of interest, using techniques like the one outlined below.

2.2 Topic Modeling and Selection of Texts

Automated selection from large volume sets is necessary because one cannot hope to inspect by eye the whole collection. Even a set of 1,315 books is on the order of Charles Darwin's entire personal library, accumulated and read over many decades. To help solve the problem of "what to read?", we have built and used a variety of tools for extracting 'latent

[4] https://inpho.cogs.indiana.edu/scimap/jisc/

meaning' factors between, and among, terms and documents.

There are several popular algorithmic approaches to the extraction of latent semantic relationships from text. These methods commonly represent terms and/or documents within a *multidimensional vector space* (i.e., a generalization of the familiar 3-dimensional physical space). The vector spaces used are typically (but not necessarily) very high-dimensional — ranging from a few hundred dimensions to tens of thousands of dimensions. Depending on the model, axes (dimensions) may correspond to single words or multiword phrases, and documents to vectors within the vector space; or the dimensions may be arbitrary, not related to anything specific, with terms and/or documents represented as vectors within the space. Again depending on the particular model, term-term similarity or term-document similarity may be computed as the cosine (or, the inverse cosine, arccos) of the angle between two vectors – or in other applications, as the Euclidean straight-line distance between the vectors.

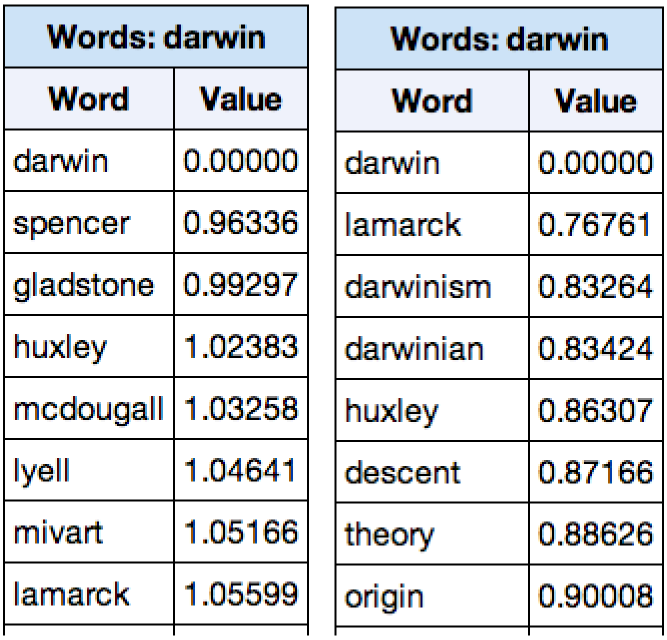
In the Digging by Debating project we implemented three kinds of vector space models (VSM) of latent meaning factors, and these are described in detail in the next section. The tools are archived in an Open Source repository.

2.2.1 Techniques for topic modeling

Three classes of vector space models were implemented and tested on the corpus: (1) standard latent semantic analysis (LSA) based on term-frequency approach known as tf-idf; (2) the BEAGLE model of Jones & Mewhort (2007) which excels at word-similarity based on sentence position and word context; and (3) topic modeling by Latent Dirichlet Allocation (LDA). Within vector space models, it is common to compute cosine values to measure similarity between vectors (or arccos if a proper metric is preferred).

The use of tf-idf (term frequency/inverse document frequency) was quickly eliminated from consideration. Its "bag of words" approach to document-document similarity is quite good where there is substantial term overlap between documents, but it is less powerful at finding meaning relationships when there exist substantive differences in vocabulary.

BEAGLE is very good at term-term similarity and has a high degree of psychological plausibility, having been tested experimentally against human subjects' semantic similarity judgments (Jones & Mewhort 2007). The model has separate word order and word context components, which can be blended to different degrees into a composite model of order and context. Different blends in composite models are may be used to represent word similarities according to semantic role (e.g., finding words that function similarly as names, nouns, or verbs; see the 50:50 blended model at left side of Table 1) or conceptual relatedness (e.g., the 70:30 blend at the right of Table 1). However, BEAGLE is less adaptable to term-document or document-document similarities. While we are convinced that BEAGLE remains a useful tool for the project as originally envisaged, for the immediate purposes of selecting books and passages for argumentative analysis in the area of comparative psychology, LDA topic modeling proved itself to be more suitable.



Words: darwin	
Word	Value
darwin	0.00000
spencer	0.96336
gladstone	0.99297
huxley	1.02383
mcdougall	1.03258
lyell	1.04641
mivart	1.05166
lamarck	1.05599

Words: darwin	
Word	Value
darwin	0.00000
lamarck	0.76761
darwinism	0.83264
darwinian	0.83424
huxley	0.86307
descent	0.87166
theory	0.88626
origin	0.90008

Table 1. *Words most similar to 'darwin' according to a BEAGLE model trained on the HTRC-1315 collection of books, according to a 50:50 order:context blend on the left and 70:30 blend on the right. The lists are ordered by the arccos of the vectors representing the terms showing how more 'order' information tends to select names, i.e. terms with a functional role similar to 'darwin', while more 'context' information tends to select for conceptually related terms (include the names of people working on similar ideas).*

LDA is a generative model which conforms to a Bayesian inference about the distributions of words in the documents being modeled. Each 'topic' in the model is a probability distribution across a set of words from the documents. As with all three methods, a stop-list is used to ignore common high frequency words. We also eliminated all words that occurred with frequency less than three, which helped to filter out the many OCR errors in the HTRC collection. The number of topics generated by the method is a user-selected parameter (K) and we have been experimenting using various values of K. Further work is needed to determine whether K can be optimized for a given task[5].

Using topic models, a word (or combination of words) can be used to identify topics, and a topic (or combination of topics) can be used to select documents whose actual word distributions are best predicted by the probability distributions within the topic(s). Because all term-document relationships are mediated by the topics, and because every topic represents, to some degree, every word in the corpus, LDA is adept and finding highly implicit relationships between documents, making it possible to identify highly abstract connections between documents in the corpus – even between documents with few, or no, overlapping words. This becomes particularly apparent in the sentence-sentence similarity examples

[5] but see Arun et al. (2010) and Heinrich (2011), for steps in this direction.

described further below in this report.

The InPhO group has made the vsm tools available via an iPython Notebook framework[6] allowing rapid prototyping and testing with the models. A considerable amount of preliminary work was necessary to coordinate with the Hathi Trust Research Center to arrange access to the volumes, and clean up of their files for ingestion to the vector space modeling. Having also worked with the Stanford Encyclopedia of Philosophy and the PhilPapers databases using these tools, we can report with reasonable confidence that corpus preparation will remain one of the most labour-intensive parts of the process due to the large diversity of formats. Also, there are no good ready-made systems for managing metadata associated with these varied corpora, nor for attaching those metadata to the products of the analyses, so there remains a large degree of special-purpose coding necessary for each new target corpus. Nevertheless, once a particular corpus has been prepared for analysis, the vector space modeling itself is highly repeatable across the different corpora, and Allen’s InPhO group is continuing to explore ways of comparing and exploring the outputs of different modeling methods.

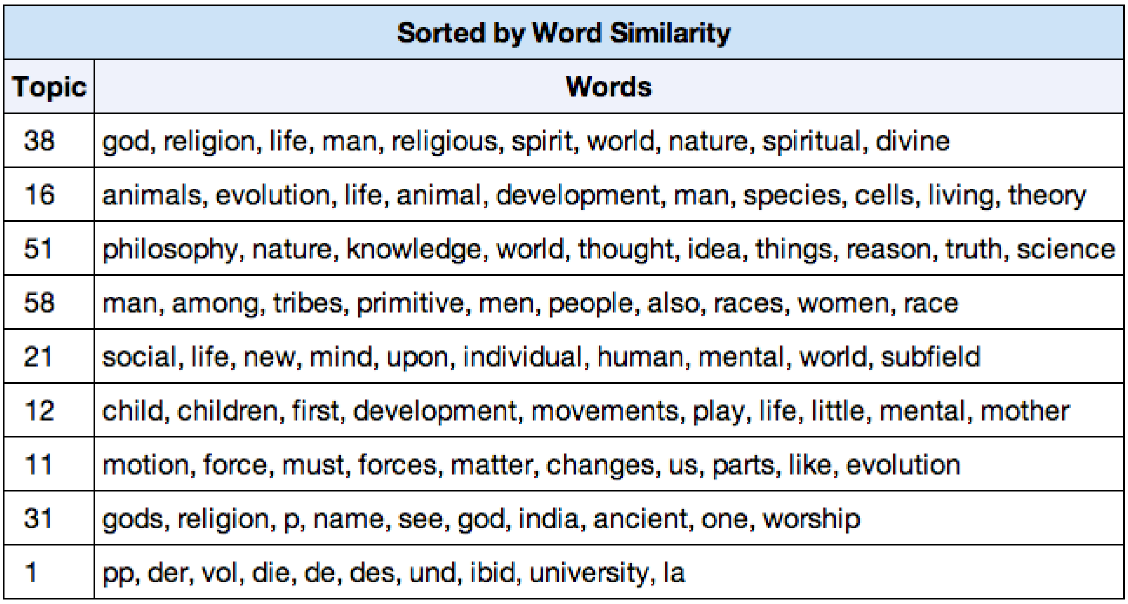
2.2.2 Selecting texts using topic modeling

In Spring 2012, with the help of HTRC staff, we conducted a keyword search in the HTRC dataset which then comprised just over 200,000 volumes published prior to 1928. Using the key words ‘darwin OR romanes OR comparative psychology’ we attempted to isolate a manageable portion of this corpus, which turned out to comprise 1,315 volumes (complete books). These books spanned a range of genres (as shown in the map, Figure 1), not just comparative psychology, but theology and various education-related titles that also included course bulletins from universities. Our goal was thus to use topic modeling to narrow even further this set to those that were specifically relevant to our pursuit of arguments about the nature of animal minds in comparative psychology.

Thus, with these 1,315 volumes as our corpus, we used LDA topic modeling to construct a set of topics. After trying various implementations, we settled on a parallelized Gibbs sampling method (Newman et al. 2009) that made it possible to run the topic models with various values of K (the number of topics) in increments of 20 from K=20 to K=140, taking a couple of days of full time processing on our 24-core Dell Poweredge T410 with 32 Gigabytes of RAM running Ubuntu linux.

Given a set of topic models it is possible to obtain topic suggestions using single words or combinations of words. For example, Table 2 shows topics most related to the word ‘anthropomorphism’ from the K=60 model (sorted by highest p-value for that topic generating the term; topic index numbers are arbitrary). The words shown on each row are the 10 highest probability words in that topic, but recall that each topic is actually a probability distribution about all the words occurring at least three times in the corpus:

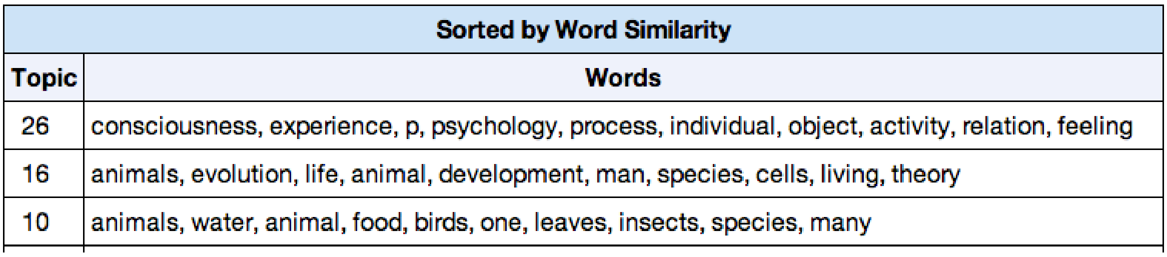
[6] http://ipython.org/ipython-doc/stable/interactive/notebook.html



Sorted by Word Similarity	
Topic	**Words**
38	god, religion, life, man, religious, spirit, world, nature, spiritual, divine
16	animals, evolution, life, animal, development, man, species, cells, living, theory
51	philosophy, nature, knowledge, world, thought, idea, things, reason, truth, science
58	man, among, tribes, primitive, men, people, also, races, women, race
21	social, life, new, mind, upon, individual, human, mental, world, subfield
12	child, children, first, development, movements, play, life, little, mental, mother
11	motion, force, must, forces, matter, changes, us, parts, like, evolution
31	gods, religion, p, name, see, god, india, ancient, one, worship
1	pp, der, vol, die, de, des, und, ibid, university, la

Table 2. *Topics most related to the word 'anthropomorphism' from the K=60 model (sorted by highest p-value for that topic generating the term)*

Inspection of this list indicates that 'anthropomorphism' relates to a theological topic (38; numerical indices are arbitrary), a biological topic (16), a philosophical topic (51), an anthropological topic (58), etc. It is also possible to narrow the search by querying topics using a combination of terms. The Table 3 shows the top three topics returned using 'anthropomorphism', 'animal', and 'psychology' as input.



Sorted by Word Similarity	
Topic	**Words**
26	consciousness, experience, p, psychology, process, individual, object, activity, relation, feeling
16	animals, evolution, life, animal, development, man, species, cells, living, theory
10	animals, water, animal, food, birds, one, leaves, insects, species, many

Table 3. *Top three topics returned using 'anthropomorphism', 'animal', and 'psychology' as input*

This technique reveals two apparently relevant topics that were not located using 'anthropomorphism' alone as the probe and we may now use all three topics to filter the documents from the original set of 1315, with the following results (partial listing only; shown here with live links to HathiTrust Digital Library):

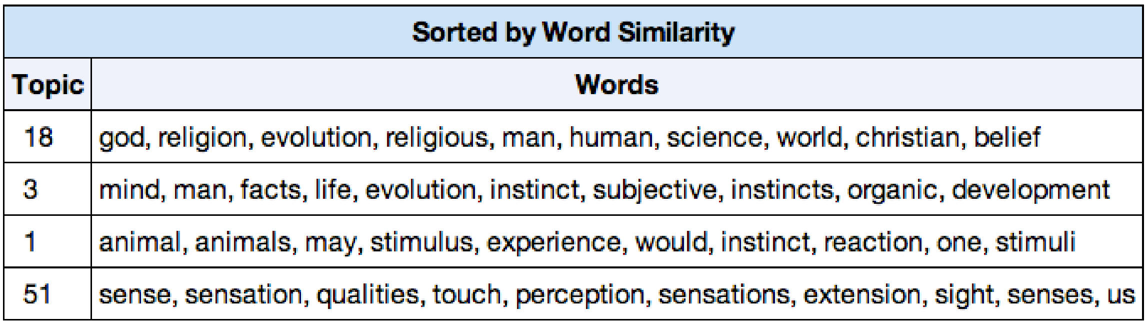
Document	Prob
Secrets of animal life,, http://hdl.handle.net/2027/uc2.ark:/13960/t7wm15g73	0.63954
Comparative studies in the psychology of ants and of higher ..., http://hdl.handle.net/2027/uc2.ark:/13960/t6057f659	0.63085
The colours of animals, their meaning and use, especially ..., http://hdl.handle.net/2027/uc2.ark:/13960/t9t14w82w	0.55333
The foundations of normal and abnormal psychology, http://hdl.handle.net/2027/loc.ark:/13960/t9m33nm99	0.54171
The bird rookeries of the Tortugas, http://hdl.handle.net/2027/uc2.ark:/13960/t3pv6cc9j	0.53789
Mind in animals, http://hdl.handle.net/2027/mdp.39015005169357	0.53783
Ants and some other insects; an inquiry into the psychic powers..., http://hdl.handle.net/2027/wu.89095158218	0.53606
Systematic science teaching ; a manual of inductive elementary..., http://hdl.handle.net/2027/uc2.ark:/13960/t11n8195t	0.53152
The riddle of the universe at the close of the nineteenth century..., http://hdl.handle.net/2027/uc2.ark:/13960/t5v69b880	0.52804

Table 4. *Results of using all three topics to filter the documents from the original set of 1,315*

2.2.3 Selecting pages using topic modeling

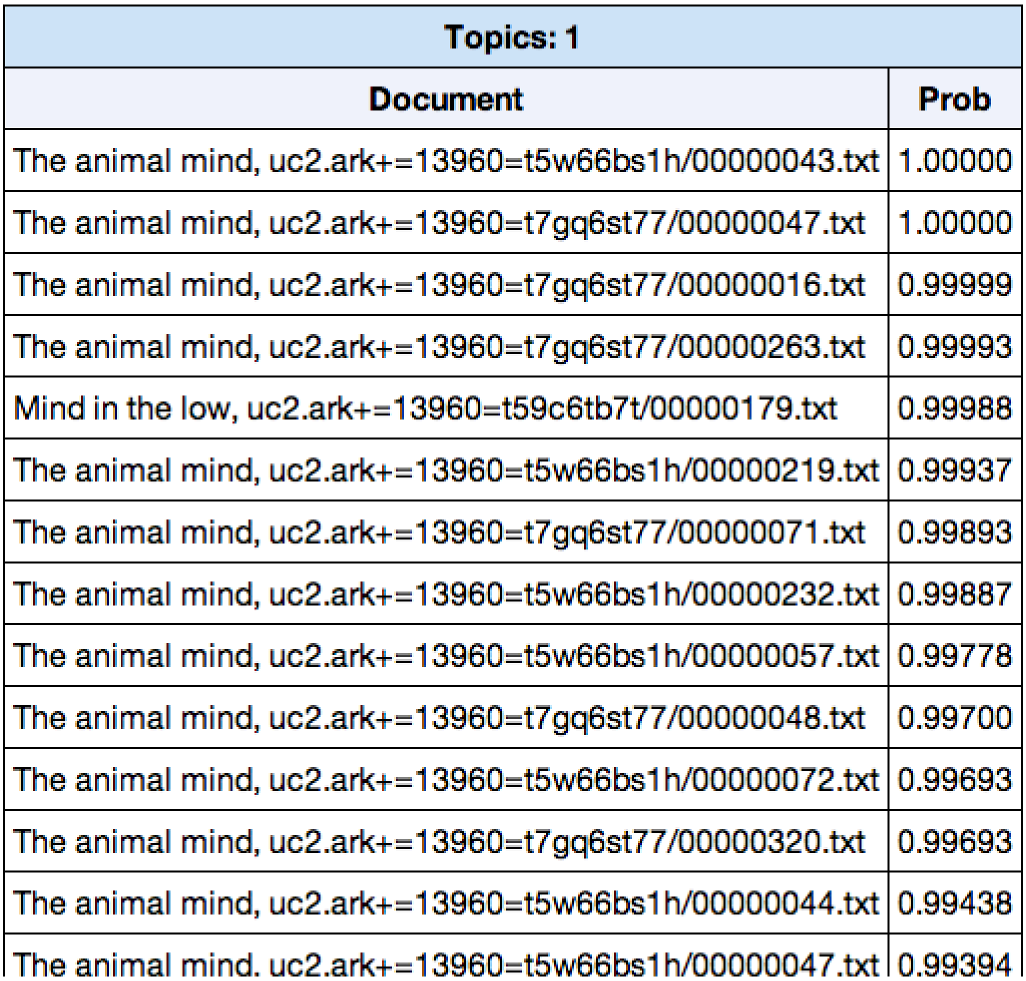
By applying a threshold on the probability values attached to these results, we were able to reduce the original set of 1,315 books to an 86 volume selection. On this sub-selection we ran a more fine-grained analysis treating each individual page as a separate document. For the sake of comparison, we again probed a K=60 topic set using ‘anthropomorphism’ as the key, with the results shown in Table 5.

Although the theological topic has not been completely eliminated from the 86-volume set, it is clear however that the biological and psychological topics have become more prevalent. Using 'anthropomorphism', 'animal' and ‘psychology’ as probes, topic 1 bubbles to the top, suggesting that for the purposes of locating specific pages relevant within the 86-volume corpus that are relevant to our initial interests, it provided the best starting point. Thus, using this topic, a selection was made of the top 800 rated pages (from the 86 volumes) with the highest p-values for the topic (Table 6).



Sorted by Word Similarity	
Topic	Words
18	god, religion, evolution, religious, man, human, science, world, christian, belief
3	mind, man, facts, life, evolution, instinct, subjective, instincts, organic, development
1	animal, animals, may, stimulus, experience, would, instinct, reaction, one, stimuli
51	sense, sensation, qualities, touch, perception, sensations, extension, sight, senses, us

Table 5. *Results of a search using a K=60 topic set using 'anthropomorphism' as key*



Topics: 1	
Document	Prob
The animal mind, uc2.ark+=13960=t5w66bs1h/00000043.txt	1.00000
The animal mind, uc2.ark+=13960=t7gq6st77/00000047.txt	1.00000
The animal mind, uc2.ark+=13960=t7gq6st77/00000016.txt	0.99999
The animal mind, uc2.ark+=13960=t7gq6st77/00000263.txt	0.99993
Mind in the low, uc2.ark+=13960=t59c6tb7t/00000179.txt	0.99988
The animal mind, uc2.ark+=13960=t5w66bs1h/00000219.txt	0.99937
The animal mind, uc2.ark+=13960=t7gq6st77/00000071.txt	0.99893
The animal mind, uc2.ark+=13960=t5w66bs1h/00000232.txt	0.99887
The animal mind, uc2.ark+=13960=t5w66bs1h/00000057.txt	0.99778
The animal mind, uc2.ark+=13960=t7gq6st77/00000048.txt	0.99700
The animal mind, uc2.ark+=13960=t5w66bs1h/00000072.txt	0.99693
The animal mind, uc2.ark+=13960=t7gq6st77/00000320.txt	0.99693
The animal mind, uc2.ark+=13960=t5w66bs1h/00000044.txt	0.99438
The animal mind. uc2.ark+=13960=t5w66bs1h/00000047.txt	0.99394

Table 6. *Partial results of a selection of the top 800 rated pages within the 86 volume set*

The p-values associated with the 800 selected pages range from 1.00 (highest probability of that page/document being generated by the topic) down to about 0.30 (still relevant).

We return to the use of LDA topic models for even finer-grained analysis – specifically, trained on individual volumes with sentences as 'documents' – in section 4 below.

2.3 Argument Identification and Analysis of Argument

In this section we report how we used semi-formal discourse analysis linked to modeling techniques to better understand the distribution and nature of arguments in the texts of interest. To move from full texts to argument models and visualisations we have previously used techniques that link discourse analysis to formal argument modeling and system design, such as the DISCOUNT scheme (Pilkington, 1999) and the Investigation by Design (IBD) methodology (Ravenscroft & Pilkington 2000) used in various dialogue game projects (reviewed in Ravenscroft, 2007). The approach builds on the use of semi-natural argumentation framework of dialogue games (Ravenscroft et al., 2009) to both structure the debates around texts and provide semantics that can be mapped to more formal argument structures that are present in philosophical texts and can be mapped to the Argument Interchange Format (AIF) developed by the Dundee team.

The outcomes of the use of semi-formal discourse analysis, linked to modeling techniques, should allow us to decide whether relevant arguments can be suitably mapped using the OVA tool (that generates arguments in AIF), and therefore support and inform additional development of automated analysis algorithms to do the same, or whether understanding and mapping arguments in our texts is particularly complicated, requiring complex human interpretation, and therefore tool development should focus on the development of visual interfaces allowing people to better understand, identify and represent arguments.

In this section we detail the process of human selection and analysis of argumentative passages from six volumes (up to a maximum of 40 pages from each). The content was marked up as a set of propositions with links to the propositions that they support or undercut. In this way each argument was mapped as a conclusion with premises that support it. Firstly, we analysed the propositions according to content, source and role. Secondly, we analysed the representativeness, type, importance, meaning and context of the whole arguments.

2.3.1 Selection of six volumes for argument analysis

We chose to analyse six volumes that are relevant to the topic chosen. We aimed to have some 20-40 rated pages from each volume as a fair reflection of the type of argument contained therein. We noted that although some rated pages scored highly, the page count from their volume was too low, and we could not be assured that these few pages are representative of the volume, so a simple metric was used to decide which volumes (and pages) should be included in this study.

The metric was to add together the associated p-values for the selected pages across each of the respective volumes to provide a ranking of volumes suited for our purpose. Although adding probability values makes no sense if one wishes to interpret the result as a probability, for our purposes it provided a ready guide to the volumes, given that we preferred on-topic content with sufficient pages to fit into our frame of analysis (20 plus pages), and cumulating the p-values of rated pages by volume gives a good indication of this. The results are shown in Table 7.

Clearly the volume 'The Animal Mind' contained a wealth of on-topic material and was the first candidate volume to be included, the next four ranked volumes having little difference between them. The final volume only just met the criteria for inclusion (20 plus rated pages).

Volume Title	Hathi code	Metric: p-values	Rated Pages
The Animal Mind	t5w66bs1h	122	154
Psychology of Ants and of Higher Animals	t6057f659	24	42
The Principles of Heredity	t74t6gs0m	20	36
General biology	t0ht2h954	20	32
The Nature & Development of Animal Intelligence	t05x26n0d	19	37
Progress of Science	t5p84550z	12	20

Table 7. *Selected volumes and metric of cumulated p-values*

2.3.2 Selection of rated pages and markup

The rating of pages was taken as an indicator of on-topic material, but not used to limit arguments that started before or ended after the rated pages. Thus, each argument selected spanned rated pages, but may also have spanned unrated pages occasionally. We should note that not all rated pages that dealt with the chosen topic contained argument. The first three volumes were argument-rich, but the last three were not, and so few arguments were mapped from the latter. For *The Animal Mind,* which contained a plethora of rated pages, to narrow down the pages to a maximum of 40 we listed those pages with p-values of 0.9 or greater, and chose to examine the largest page blocks (contiguous pages). For the six volumes, the pages containing argument and actually used in our PassA are shown in Table 8.

The argument content was marked up in OVA+, an application which links blocks of text using argument nodes. OVA+[7] provides a drag-and-drop interface for analysing textual arguments. It is reminiscent of a simplified Araucaria[8] except that it is designed to work in an

[7] OVA+ (http://ova.computing.dundee.ac.uk/plus) see also Lawrence et al. (2012)

[8] Araucaria (http://araucaria.computing.dundee.ac.uk/)

online environment, running as a Flash widget in a browser. It also natively handles AIF[9] structures. Each argument, as selected in the previous section, was divided into propositions and marked up as a set of text blocks. These text blocks containing propositions were linked to propositions that they support, or undercut, to create an argument map.

Volume	***Maps***	***Pages***
The Animal Mind	15	13-16, 16-21, 24-27, 28-31, 31-34, 58-64, 204-207, 288-294, total = 40 pages (original page numbering)
The Psychology of Ants	10	Preface, 15-19, 31-34, 48-53, 99-103, 108-112, 206-209, 209-214, total = 37 pages (renumbered[10])
The Principles of Heredity	8	374, 381, 382, 385, 386, 390, 394, 395, total 10 pages (renumbered)
General Biology	2	434-435, 436 total = 3 pages (original page numbering)
The Nature & Development of Animal Intelligence	5	16-18, 21-26, 30-32 total = 12 pages (renumbered)
Progress of Science	3	479-484, total = 6 pages (renumbered)

***Table 8**. Page lists of analysed pages from selected volumes*

There are 60 OVA+ maps for argument mark-up which can be viewed online[11]. An example is included in Figure 3. Besides PNG files, Zips of the JSON files are available which can be loaded into OVA+ (showing exactly how the plain text of the volume was marked up), and a document containing the plain text of the whole volume. The links drawn on the maps between propositions are of two types - supporting and undercutting (on the link seen as RA and CA, OVA+ supports more link-types but these are not used in this study). Conclusions must be supported by at least one premise. Often the maps have sub-conclusions leading to main conclusions. Propositions that expand or explain other propositions are seen as lending support to them. In theory, it is possible to have two propositions that support each other, but it is confusing to read and is avoided. A link connecting two propositions always links from one to another, with an arrow showing direction, where a supporting premise links to (points to) a conclusion or sub-conclusion.

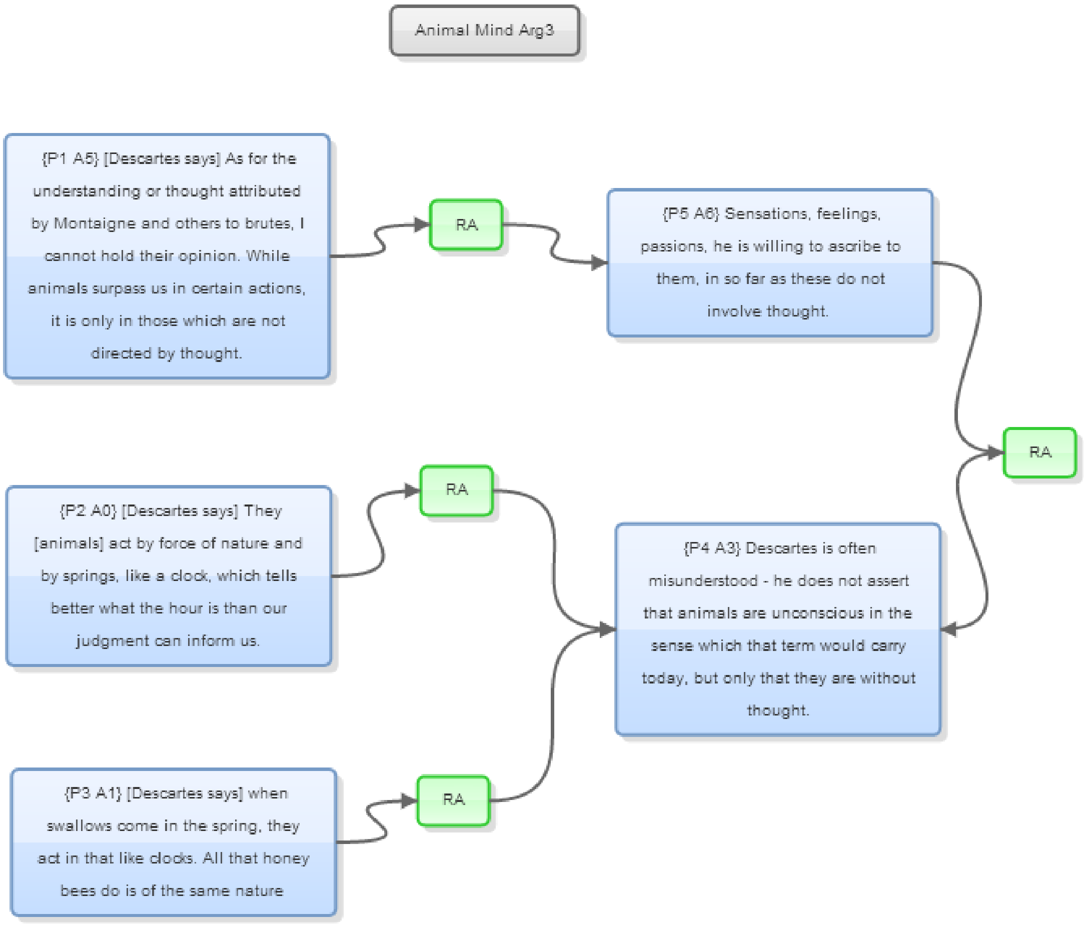
[9] AIF (http://www.arg.computing.dundee.ac.uk/?page_id=197) see also Chesnevar et al. (2006)

[10] Original page numbering masked by a bug

[11] Online link (http://bit.ly/1bwJwF9) to volumeData on Google Drive (view only). Open volume folder, open the pass subfolder, select a PNG image and see a Preview pane - click the blue OPEN button. A new tab will open – click button 100% to zoom. Move around by dragging the diagram.

Arguments can be mapped at different levels depending upon the choices the mapper prioritises. This is particularly true of volumes such as are analysed here, where, in some cases, the same topic is pursued for a complete chapter and so there are opportunities to map the extended argument. The first two volumes were marked up at two levels, while the last four were marked up with only the lower level.

Pass 1 (17 maps): the objective in this mark-up was to identify a simple but coherent ‘thought process’, abstracted from the vagaries of how it is expressed. In most cases the text block has been paraphrased - so that the node display expresses the ‘thought’ in a concise and direct way. Pass1 maps the arguments in summary form with each map containing several specific arguments (and possibly counter-arguments).



Animal Mind Arg3
{P1 A5} [Descartes says] As for the understanding or thought attributed by Montaigne and others to brutes, I cannot hold their opinion. While animals surpass us in certain actions, it is only in those which are not directed by thought.
RA
{P5 A6} Sensations, feelings, passions, he is willing to ascribe to them, in so far as these do not involve thought.
RA
{P2 A0} [Descartes says] They [animals] act by force of nature and by springs, like a clock, which tells better what the hour is than our judgment can inform us.
RA
{P4 A3} Descartes is often misunderstood - he does not assert that animals are unconscious in the sense which that term would carry today, but only that they are without thought.
{P3 A1} [Descartes says] when swallows come in the spring, they act in that like clocks. All that honey bees do is of the same nature
RA

Figure 3. *Sample argument map from OVA+*

Pass A (43 maps): the maps each contain more detailed argument, and propositions have been marked up with Representativeness codes from the Analysis scheme. Each text box is labelled with a proposition number in the order that they occurred in the original text. Where there are quotations it is made clear who used the words with the actual author named if necessary in square brackets, (for example, as [Descartes says]) at the start of a text box.

It is possible to mark-up content at different levels of detail, depending upon the purpose required. The most detailed maps would contain one conclusion and one or more premises, whilst summary maps (like Pass 1) may contain several distinct, but related, arguments on one map. Pass A here, is not the most detailed of maps, since it contains sub-conclusions, but these build a substantive argument rather than fragmentary ones.

2.3.3 Scheme for analysis of argument

There are three elements to the analysis of argument scheme presented here -

1) **Representativeness** of the display text compared to the original text. The original text may be verbose and there is scope for reducing and simplifying the text, whilst keeping the essential meaning – codes are posted on each text box in the maps (note that this element is not strictly analysis, but interpretation).

2) **Types of proposition** and the role they play within the argument structure. We want to record what types of proposition are used and the occurrence of explanation, evidence, reference to other authority etc. Unstated premises should also be noted.

3) **Importance, meaning and context** of the argument. The main topic is Animal Consciousness and Anthropomorphism. Arguments that deal with these topics are marked as important. We are also interested in the meaning of arguments, are there multiple meanings, and the part that context plays in enhancing meaning for the reader.

The Representativeness codes are one letter and one digit posted within { } (after the proposition number) at the start of a OVA text box, as follows -

- A - Actual words (apart from minor grammatical changes)
- M - some parts have been Modified, but most text is from original
- R - Rewritten text, most of the text shown has been summarised
- U - Unstated in the original, make explicit the implication of text

The digit represents the amount of original text excised from the text box (as a percentage divided by 10). So 0 is 0-9% removed, 1 is 10-19%, 2 is 20-29%, etc. With code 0, no changes or only minor grammatical changes have been made. Examples are

- {P5 A0} - proposition 5 has original text, nothing removed
- {P2 M5} - proposition 2 has had 50% of original text cut and some words modified
- {P1 R9} - proposition 1 has been rewritten and contains almost no original text

The second element of the scheme is to examine proposition types and the role they play within the argument. The term proposition is used to cover both premises and conclusions. We also noted the possible syntactic signifiers that are present in the original text of each proposition that might be useful in machine analysis of argument (shown in analysis worksheet).

		Proposition Content
X	Example	An example to demonstrate point
O	Observation	Detailed observation of animal/human behaviour as evidence
V	Evidence	Other evidence, from fact or authority, to back up assertion
E	Explanation	Explanation or expansion of earlier point
R	Reasoning	Well-grounded reasoning
C	Criticism	Criticism of another's argument
A	Assertion	Simple assertion - other than above

		Proposition Source
G	General	General reference to others' arguments
R	Reference	Specific reference to authoritative writing, own or others
Q	Quotation	Quotation of authoritative writing
O	Observation	Detailed observation by author
N	Anecdote	Anecdote, hearsay or non-authoritative writing, and not based on detailed observation
A	Assertion	Simple assertion - other than above
U	Unstated	Unstated/implied supporting premise

		Proposition Role
C	Conclusion	Main conclusion, other propositions link to it, it links to no other
B	Subconclusion	A conclusion, propositions link to it, but it links to another
S	Support	Supporting premise, links to a sub/conclusion
X	Counter	Counter argument, link with CA to sub/conclusion

Q	Qualifier	Premise with qualifier to scope of argument
O	Opposite	A statement of position, where the mapped argument is the counter to this

Table 9. *Proposition analysis schemes, 3 facets – content, source, role*

Each of the propositions in the mapped arguments were assigned three codes as in Table 9, to represent the three facets. The coding was implemented so that each code is (as far as possible) mutually exclusive within the category. Note that, on category Q, the decision was taken to include any qualifying statements in with a premise, since they do not either support or undercut the conclusion - and we only have two types of links in the maps.

The third element of the scheme is more discursive - to examine the importance and meaning of the argument within the text of volume analysed. Is this argument a stepping stone for other arguments? What part does context play in stating the argument, and what part in interpreting meaning from the argument? Is the argument strong, or are there weaknesses? Is the argument clear?

3 Results: Analysis of Rated Pages from Six Volumes

This section presents the results of the (human) analysis of selected argumentative text in the six volumes. We use the quantitative analysis scheme presented in section 2.3.3 to categorise the propositions used in the arguments according to content, source and role. A qualitative analysis of arguments, their importance, meaning and context then follows.

3.1 Quantitative Analysis of Propositions

A simple Chi-Squared analysis[12] of the counts of the proposition types in each category by volume, in the scheme presented previously, showed that there were highly significant differences between some volumes for Propositional Content and Propositional Source as the table of probability values shows (Table 10) .

The greatest differences noted under Propositional Content are the extensive use of examples (category X) in the Heredity volume versus the lack of examples in the Ants volume and the extensive use of detailed observations (category O) in the Ants volume compared with the lack of detailed observations in the other two volumes. Another large difference is the greater use of evidential statements in the Animal Mind volume (category V).

[12] The worksheet for analysis of propositions can be found on this link http://bit.ly/1fj38y7

Volume	***Content***	***Source***	***Role***
The Animal Mind	0.1793	**0.0000**	0.4537
Psychology of Ants and of Higher Animals	**0.0000**	**0.0000**	0.9430
The Principles of Heredity	**0.0000**	**0.0211**	0.7435
General biology	0.3438	0.5719	0.8830
The Nature & Development of Animal Intelligence	0.1818	**0.0393**	0.1039
Progress of Science	**0.0021**	0.4886	0.7070

Table 10. *P values of each propositional facet[13] with significant values bolded.*

The greatest differences under Source are the extensive use of references (category R) and quotes (category Q) for Animal Mind compared to the other two volumes. By contrast, there is extensive use of detailed observations of animal (ant) behaviour by the author in the volume Ants and none in Animal Mind (category O). Finally, there is slightly more use of simple assertions by the author in Heredity (category A). Under the Role category there are not great differences, apart from slightly more qualification (category Q) in the arguments of Animal Mind.

These differences are commented on in the next section[14]. Moving on to the third element in the analysis scheme, we are able to discuss the meaning, importance and context of the arguments mapped in Pass A of each of the three volumes. Throughout this discussion, ‘animal’ is to be taken as not including Man.

3.2 Qualitative Analysis of Rated Arguments from Six volumes

3.2.1 Volume 1 analysis - The Animal Mind, 1908

The author, Margaret Washburn, sets the context for the debate on animal consciousness in which there are a wide diversity of views. She meets head on one of the fundamental criticisms used to dismiss animal psychology straight away – that all psychology must be anthropomorphic – and admits there is a problem (Arg1). She introduces the arguments of Montaigne for the similarity of animal behaviour with Man's. Montaigne was a humanist observer (of man and nature) and renaissance writer of the 16th century, with a natural

[13] Null hypothesis: counts of proposition types within category in each volume do not differ from expected values - the table shows the probability in each category (bolded values are significantly different from the null hypothesis, compared with expected values as taken from all six volumes (see link Analysis). No correction made for small cell-values)

[14] In fact the comments were originally written without the benefit of the analysis, but the quantitative analysis corresponds with the qualitative comments.

sympathy for animal intelligence and recognition of animal behaviours (Arg2). The response by Descartes, a mathematician and philosopher, who shortly followed Montaigne is that animals are like automata – though with feelings (he also likened the human body to a machine). In his view, animals have an assigned place in the world, which cannot include 'thought', reserved by God for Man only. Although Descartes' references to swallows and bees are mentioned by Washburn, she emphasizes his view of their behavior as clock-like (Arg3).

Darwin's argument comes out in a rather weak form since there is no illustration of the points made. Since animals dream they must have imagination. They are 'seen to pause, deliberate and resolve', so they have some reasoning. A twist here is that the author posits an ulterior motive to Darwin's writings which are aimed at the ongoing controversy about the mental and moral gulf between man and animals. Thus, the author warns, not all can be taken at face value (Arg4). In contrast are the arguments of many physiologists, who do not allow a psychic interpretation of the animal mind. They prefer to think of biological explanations (like tropism - an unconscious reaction to stimulation) for all behaviours (Arg5).

The next argument is included for clarity. It judges that there are three main camps or positions in the field, and which writers sit within which camp (Arg6). The author sets about the physiologists' arguments, showing how they ignore or simplify in order to fit a predetermined theory. Using their approach we have the reductio ad absurdum argument when it is applied to human behaviour (Arg7). The arguments of Eric Wasmann (the next author in our volume set) for the variability and modifiability of animal behaviour are tested next from the second camp of ideas. Wasmann has extensively studied ants, and his writings were of world renown (and not just among experts). His definition of intelligence explicitly excludes animals because, he says, they can only act on instinct without deliberation. He generalises from ants to all animals, and he states ants are superior to other animals, though not apparently intelligent. It appears that Wasmann is torn between his evidence of animal behaviour as intelligent, and his theistic world view (he is a Jesuit) in which it is necessary to set Man apart from other animals, hence his definition that only Man can have intelligence (Arg8).

Washburn presents the arguments of Bethe from the third camp, part of an ultra-Cartesian group who reject all that is non-material, and who would prefer not to believe that animals have sensation. There is some inconsistency in his writing (according to Washburn), when he acknowledges modifiability of behaviour as an indicator of consciousness, but in later writing considers this an improper assumption if applied to animals. He damns all psychology as subjective and unknowable, and believes only chemical and physiological processes should be the object of scientific investigation (Arg9 and Arg10).

The author presents her own views and posits a cautious, but rational and scientific, approach to animal psychology, which acknowledges pitfalls and problems, but seeks to adjust method to overcome them, and cautions against regarding human and animal conscious processes as identical (Arg11). She introduces a conservative principle of interpretation from Lloyd Morgan (Morgan's Canon) to reduce anthropomorphism – whereby the simplest level of psychic faculty for an animal should be assumed that can fully explain the facts of a case. The author points out that the choice may not always be the right one, but at least it compensates for a known bias, and that is the best that can be done without further information (Arg12). The author asks if 'learning by experience' is the criteria of mind in animals. Loeb suggests if an animal can be trained, it has mind. The author argues that 'learning by experience' cannot be a conclusive criteria for proof of mind, and that absence of proof for consciousness does not amount to disproof. Rapid learning practically assures proof of mind, but the author avers that great uncertainty remains about consciousness in lower animals (Arg13 and Arg14). Morphology and similarity of animals' physiology to humans' must be taken into account in deciding if an animal is conscious or not. The author introduces the matter of degree of similarity, indicating a graduation of consciousness, from lower to higher animals. She is clear that nobody is able to draw a line between animals with and without consciousness (Arg15).

Nearly all the arguments above fall into the 'Important' category - dealing with animal sensation and consciousness. These arguments take the reader through a major controversy of the time regarding animal consciousness, and each of the various camps has their arguments presented and thoroughly critiqued, sometimes directly, sometimes indirectly. The writing is evidence-based and therefore scientific, albeit with ponderous phrasing that would not be used today, and has clear notions of acceptable evidence and proof.

3.2.2 Volume 2 analysis - Psychology of Ants, 1905

The author Eric Wasmann dedicated his life to the study of ants and was renowned for revealing to the general public the variety of (amazing) ant behaviours. The intelligence of animals is only instinct because “intelligence is a spiritual power” and if animals had this spiritual power “they would necessarily be capable of language”, he asserts. Animals don’t speak, so animals don’t have intelligence (Arg1). The author's position has assigned Man ‘a priori’ to a different level due to this spiritual power, and justified it by reference to (especially) language ability. The author has much evidence to expound on 'intelligent ant behaviours', but has taken care to keep this ant intelligence separate from (human) intelligence. He notes observations made by Aristotle, Stagirite, St Augustine and a contemporary naturalist, Dubois-Reymond, on the wonder of ants above other animals (Arg3). None of these authors would appear to be oriented towards science however, as Wasmann appears to be, so the choice is odd.

He denigrates any suggestion by ‘modern sociologists’ that ant states and human republics can be equated, and explains how class differences arise from ‘conditions of life’ or ‘intelligent’ free choice in Man, but ant castes arise from organic laws of polymorphism – where the body may morph into one of several types (Arg4). And, whereas “ants have perfect equality and fraternity”, “Man is egotistical and often prefers his individual welfare to the common weal”. It appears that Man’s intellect, education and customs conflict with social instincts, and while communism is good for ants it will never do for Man.

Wasmann asserts animal intelligence is really sensile cognition and sensous experience, but if higher animals are credited with intelligence, it would be inconsistent to deny ants the same (Arg5). Ants actuation of the social instincts is guided by sensile cognition, and they achieve the highest degree of community life, so they must be credited with the “highest degree of animal intelligence”. Higher vertebrates, apes etc., bind together through social instincts which enable them to co-operate for the purposes of procuring food and mutual defence. Since ants do the same, but of a ‘more perfect nature’, they must vie with the higher vertebrates (Arg7). The argument illustrates the great independence and variableness of actions of ants in their activities, which include “herding, hunting, agriculture and horticulture”.

Wasmann criticises Darwin as anthropomorphic. In taking Darwin to task, he somewhat exaggerates the case by turning Darwin’s phrase ‘silence and obedience’ regarding a group of baboons, into ‘fidelity and obedience’, which in turn implies ‘reasonable, voluntary subjection to the demands of duty and authority’. This is somewhat easier to ridicule than the original phrase. Wasmann believes the more likely explanation for baboon behaviour is “the instinctive association of certain sensile perceptions with certain sensile impulses” (Arg6). This association removes the need to allow animals thought (a Cartesian view). He denies that intelligence need be attributed to animals at all, and that instinct is a sufficient explanation (Arg10). The author explains that instinct has two elements, ’automatism’ of behaviour (generally found in lower orders of animals) and ‘plasticity’ of behaviour (generally found in higher orders).

On the architecture of ant’s nests, which varies from species to species, very different architectures occur, even when the physical attributes of the ant species are highly similar, so a simple explanation of the variety of architecture linked to physical attributes will not do. He says the decisive factor is the psychic disposition of the ant species (Arg8), though there are a few species whose certain physical attributes lead them to specialise in certain architectures.

Wasmann considers care of the young animal as “always directed by sensitive impulses and perceptions” and ants ‘verge on heroic unselfishness’ towards their charges. Man, however, is conscious of duty and the morals of parental love. Although he admits that some of motherly love is based on instincts, motherly love cannot be attributed to animals because it is ‘spiritual’ based on awareness of duty, which only humans have (Arg9). Again, the author makes assertions about common behaviours in which humans are attributed some spirituality and therefore animals’ behaviour, although apparently similar, must be of a different cause.

The main thrust of the writing appears to be scientific when the author is writing on the subject of ants, but often veers into polemic and unsupported assumption when issues of intelligence or culture in the animal world are broached. Half of the arguments presented here fall into our 'important' category. In summary, Wasmann lists the achievements of ants and at every opportunity compares them favourably with higher vertebrates and their achievements. What Wasmann does not do, which weakens his case somewhat, is provide any specific examples of higher vertebrates activities and social instincts. One is led to the supposition that he knows little about higher vertebrates and cares less.

A contradiction runs through the arguments in which he insists that so-called animal intelligence is just ‘sensile cognition’ and that only Man has true intelligence. Yet, he also describes complex ant activities, including solving problems and based on learning from experience, that would undoubtedly, if transcribed to the human sphere, be called intelligent. Wasmann, does not accept Evolution theory and assigns a spiritual element only to Man. He is therefore saying indirectly that Man is set apart from animals by God and that animals cannot share certain qualities. He associates intelligence with this spiritual element (he does not say why), and therefore, by definition, intelligence can only be possessed by Man.

3.2.3 Volume 3 analysis - The Principles of Heredity, 1906

The author, Archdall Reid, makes a distinction between instinctive impulse, which is inborn, and rational impulse, which is acquired. The argument illustrates how instinctive activities in lower animals are achieved with inbuilt abilities which do not improve with use and experience. However, the rational activity of Man proceeds from an imagined goal using his memory and solves the problem in one of many possible ways. Unlike the instinctive creature, Man has no specific natural dexterity for tasks but shows improvement through use and experience (Arg1).

The author speculates that, in all probability. instinctive behaviours in Man and animals, aroused by a given stimulus, are facilitated by emotions. Therefore, emotions of sexual love or parental love may occur in a wide range of animals, as well as Man (Arg2). He makes a distinction between a ‘mindless’ reflex action and a ‘mindful’ instinctive action. Indeed, some instinctive actions will require a highly developed mind to accomplish them. The emotion marks an advanced stage of evolution and provides the driving force for an animal to achieve the more complex instinctive actions (e.g. spinning a web or seeking prey) (Arg3).

Reid describes how the well-tended infant is provided with a much less hostile environment in which it can begin to make ‘acquirements’. It appears that intelligence is inversely proportional to instinct in ‘higher’ animals, and that this largely (but not entirely) corresponds to the helplessness of newborn infants. Whereas, those animals whose offspring immediately perform instinctive behaviours, are lower in the ‘animal intelligence’ scale (Arg4). He asserts that social creatures can make ‘acquirements’ (or learning) more easily, and provides the example of ants as social creatures. Young ants ‘play’ and are ‘taught duties’ by older ants. Slave ants, from raids on other nests, are always raised from pupae and perform lifelong duties in the nest (Arg5). One species relies entirely on slave ants and can do no work themselves. It is not clear from the argument if this is an acquired trait or one that is inbuilt (instinct).

Reid uses a hypothetical argument to demonstrate differences in development between lower and higher animals, by comparing a frog and a human kept alive until maturity in a confined space. He argues that the frog when released would be ready to take its place in the normal environment without disadvantage since all its behaviour is instinctual, whereas the human would be a physical and mental degenerate since it requires physical and mental stimulation to achieve normality. He concludes that environmental factors greatly affect a human being’s acquirements and the outcome of maturation in humans (Arg6).

Reid defines reason as the general power to draw inferences, which depends upon memory. He argues that those animals able to make acquirements must have memory, and hence be able to make inferences from them. He accepts this is a ‘loose’ definition and that others have stricter definitions of the power of reason (he recognises that the definition is important). He argues that Man’s ability to reason is only different in degree from the oyster learning to close its shell. He concludes the human baby moves from lack of reason as it develops, to ‘recapitulate the life-history of its race’ and develop the higher levels of abstract thought, it represents the manner of its evolution (Arg7). This is a contentious conclusion since, although oyster and baby share the same lack of reason, there is no direct parallel between species evolution and the development of individual humans.

The author argues that instinct has largely been supplanted by reason in Man. Why should this be so? He argues that firstly, reason confers adaptability, which is a great advantage. Secondly, that the number of instincts that Natural Selection is able to foster at one time is limited, since it requires elimination of individuals without that instinct, and too many stringent eliminations will eradicate the species. Reason enables substitution for any number of instincts allowing Man and animals to adapt and proceed to a higher evolution (Arg8).

The author appears to write for a well-informed, but not specialist, audience. He convincingly works through his arguments using thought experiments, or well-chosen examples, and refers to other authorities only occasionally when they have convincing evidence. He is an advocate of Evolution theory and sees animals and Man on a continuum of abilities, or attributes.

3.2.4 Volume 4 analysis – General Biology, 1910

The author, entomologist James Needham, is writing a text for students of Biology and therefore the writing tends to the factual and descriptive. Only two arguments are mapped from the rated pages on this volume, because of the primarily descriptive rather than argumentative nature of the text. The first argument concerns anthropomorphic attributions particularly of emotions and feelings (hunger, satiety, anger, fear, pain, pleasure, but also curiosity) to animals. Needham traces the problem of such attributions to the generic problem of other minds, and “the fact that the mind is directly accessible only to its possessor”. The second argument advocates for a developmental approach to “the psychic life … of the lower animals” as a means to better understanding of animals and towards the improvement of human education. There are no references to other writers or alternative positions in these pages, but there are clear strands of ideas that would also be present in the work of the John B. Watson, whose behaviorist manifesto would be published in 1913.

3.2.5 Volume 5 analysis – The Nature & Development of Animal Intelligence, 1888

The author, Wesley Mills, starts by noticing that Man has sympathy for others, both men and animals, particularly those that appear to share our feelings. Religious teaching, which elevates Man above animals, tends to eliminate these sympathies in adults who regard animal life as inferior, yet much of life in ways of thinking, feeling and acting are in fact shared with higher animals (Arg1). He recommends studying the development of animals to gain a comprehensive view of the psychic life of animals. Comparative study of the development of different species and the effects of differences in environment are useful, perhaps essential, to understanding animals, and to some extent mankind (Arg2).

Mills notes that evaluations and definitions of animal intelligence vary greatly across experts, and takes some to task for their belittling of animal behaviours. He points out that, for the vast majority of humans, their day to day mental life does not employ higher intellectual processes – although they have the possibility to use them (Arg3). He rejects as simplistic bias Morgan's Canon (mentioned in the first volume), which interprets animal behaviour as caused by the simplest process concordant with the facts. He suggests the same simplistic bias is present when human beings judge their fellows hastily and imperfectly, when in fact motives are more complex (Arg4).

Mills makes the argument that although there is a vast difference between the highly educated man and the higher animals, this gap is not so great with the lowest developed man, and that psychic life differs from animals only in degree, not kind. He attacks the view that Man is the centre of the universe and for whom all other life forms exist to serve. Finally, he raises the notion of animal rights that must be recognised once Man is acknowledged to come from the wider animal world (Arg5).

The author makes important arguments, particularly with regard to judgements made with insufficient evidence, or those that serve a particular position. His early championing of animal rights is novel.

3.2.6 Volume 6 analysis – Progress of Science, 1908

The author, Arthur Thompson, disparages the two extreme positions on animal intelligence. For those that see “the man in the beast” he rebukes as childish ('at the feet of Uncle Remus') and anthropomorphic. There are those at the other extreme, who “erect between themselves and the beast a high wall” and are not curious enough to look over it and notice “Not all man is in the animal, but all animal is in the man” (Scheitlin) (Arg1).

Thompson admits consciousness in the higher animals, but for lower animals (he includes ants) it is a matter of opinion. Though much of animal behaviour can be explained by reflexes it does not prove the absence of consciousness. Unless we can get close enough to detect individual peculiarities which reveal intelligence, we should hold judgement, as much for animals as for fellow men. There are enough examples of memory, profiting by experience, and departing from usual reflexes among lower animals that a psychic description is unavoidable. And if we add the structural similarity of the nervous system of animals to our own, we are unable to accept the physiologists' arguments that they are “reflex machines” (Arg2). For the vast majority of animals (above sea-anemones, jelly-fish and worms) the physiologists view of reflexes cannot account for behaviours, and even if they were able to account for one more animal, the net result is only a reduction in the number of scientific formulae, that is all (Arg3). By this the author consigns the physiologists views of animals as “reflex automata” to only the lowest animals, allowing a concept of mind in the majority of animals. The arguments made are all important and represent (as does the whole volume) the mainstream of science at the time.

3.3 Summary of Qualitative Analysis

3.3.1 The writing style and positions of the authors

It is noticeable that much of the controversy in *Animal Mind* revolves around whether Man has a special position, separate from the animal world, and perhaps given by God. This is spite of Darwin's work, published nearly 50 years before (the full import of Darwin's work was only accepted in the 1930's). Another point of contention is in use of words. 'Thought' and 'intelligence' are used by various authors in very restricted ways, without precise definition (that we know of), which lead to widely differing conclusions.

The author of *Animal Mind*, Margaret Floy Washburn, writes clearly in a measured and cautious way, which contrasts with the incautious approach of another of the quoted authors. The arguments are well presented, with extensive quoting, and criticisms or qualifications made afterwards. Sometimes there are hints as to context - for instance in the reference to 'the Jesuit Wasmann', which inform the reader that Wasmann will be bound to a theistic viewpoint. The author's position is Evolutionist, in that Man is on a continuum with animals, and that no Deity is required to explain Man's achievements.

The style of writing in *Animal Mind* is scientific - with cautious and evidenced claims, deficiencies in knowledge pointed out and the possibilities that may be the case acknowledged - so it withstands modern-day scrutiny surprisingly well. The author, as a woman writing a scientific treatise, would have been unusual at the time, and no doubt some of the caution in her writing may stem from this. The author goes on to argue the case for Comparative Psychology, which, while it has limitations and difficulties, is best suited to address questions in this field.

The second volume, *The Psychology of Ants* by Eric Wasmann (the author identified by Washburn as a Jesuit), is in contrast to the first. The writer is obviously a world authority on ants, at a time when there was much interest in their newly revealed secrets to the public. However, his grasp on the psychology of the rest of the animal kingdom appears to be nothing less than slight. In spite of this, he feels able to make pronouncements on important issues like animal intelligence on the basis that ants have the highest social instinct and their society is 'more perfect' than any other, so what applies to ants must apply to all other animals too. The writer shows no caution in his statements about the whole animal kingdom and does not provide evidence of any kind in this selection, except in so far as it applies to ants.

This writer is still performing a philosophical (pre-scientific) role in his writing, of adopting a position first and proposing hypotheses which agree with his clearly anti-evolutionist stance. A more scientific approach would be to examine the evidence available and see what it suggests, and then to formulate a hypothesis that will explain it. The writer makes no attempt to examine evidence for animal intelligence other than that for ants. Instead, he makes assumptions and generalisations which are unjustified, but necessary in order that his world view is maintained. He is rigorous about rooting out the anthropomorphism of Darwin's (and others) writing, and insists on using only instinct (sensile cognition) to explain all animal behaviours. However, he is quite willing to inject a spiritual dimension into Man's thinking when compared with animals.

The third volume, *The Principles of Heredity* by G Archdall Reid, is more similar to the first, in that the author takes an Evolutionist perspective and presents arguments clearly. He writes with less caution than the first author, and with less caveats. It seems probable his intended audience is the well informed reader, rather than the student of Comparative Psychology, and he uses less references to others in the field and more hypothetical examples in which he demonstrates his point.

His examples are well chosen vehicles for reasoning about abilities and behaviours. The comparison between a caterpillar building a cocoon and a man building a house demonstrate the difference between instinctive and 'rational' behaviours. The hypothetical example of a frog imprisoned without stimuli and a human imprisoned without stimuli, clearly demonstrate the reliance of the human species on 'socialisation' and 'environment' (not so the frog). His clear use of examples such as these enable the general reader to understand the concepts and major conclusions to be drawn, without the hedging around what is not presently known of the first volume.

The fourth volume, *General Biology* by James Needham, is written in standard textbook style of the time. It contains few arguments and it is written in a factual scientific style, where the science is known. The author spends some time discussing the evolution of Man in the latter part of the book.

The fifth volume, *The Nature & Development of Animal Intelligence* by Wesley Mills, is a collection of addresses (speeches) by the author on the nature and scope of Comparative Psychology in the first part. In the second part are studies, documented facts about animals and their behaviour, and extracts of discussions had with respected experts in the field. Based upon his own varied and detailed observations of the development and behaviour of common animals he holds a position that animals have talents and abilities that are not yet properly documented which reveal that they are psychically closer to, yet different from, Man than most people would give credit. He is a naturalist with a scientific approach to understanding animal behaviour in its natural state, rather than in a laboratory, and so he runs a difficult line between use of documented facts and anecdote. Nevertheless, he manages to document many and varied instances of the development and behaviour of dogs, pigs, horses, pigeons among others, from his own careful observations and constructed tests of their abilities. He appears to assess evidence from third parties as to its likelihood and reliability, comparing it with what he already knows, and so avoids the pitfalls of anecdote and anthropomorphism.

3.3.2 The key arguments of the texts

A number of themes that are historically interesting emerge from these six texts, and the ability of our selection methods to allow a reader (SRM), who was previously unfamiliar with the scholarship in this area, to zero in on the relevant passages is a measure of the success of those methods. Although it was not a primary goal of the Digging by Debating project to produce new insights into the domain-specific content, these will be analyzed in future work.

Nevertheless, it is worth making some content-specific remarks here.

Several themes emerge across the six volumes. We will mention two:

(i) the authors, evolutionists and non-evolutionists alike, all are willing to recognize hitherto unacknowledged flexibility and variability in behavior of individual animals; and

(ii) the significance of a developmental approach to evolutionary studies that animal studies (but not human) studies can facilitate.

Flexibility and variability

All the authors sampled identify two extremes of excessive anthropomorphism ("man in the beast" in Thompson's (1908) memorable phrase; "Progress Arg1") on the one hand, and on the other hand, the conception of animals as automatic reflex machines (cf. "blind instinct"; Mills 1888, "Nature Arg1"). However all these authors claim the middle ground. Even Wasmann, the lone anti-evolutionist in our sample, denies that individual ants are reflex machines, claiming in the argument beginning p.112 that the individual flexibility of ants is of a "psychic variety" not "mechanical automatism". Wasmann, however, attributes this flexibility to "instinct" and is inclined to play up the flexibility of ants claiming them superior to baboons in several respects (rhetorically challenging the presumed evolutionists' idea of the gradual emergence of more and more sophisticated mental capacities). Wasmann claims that in all cases, animal flexibility, being based in instinct, has a wholly different source than the human "spiritual power of abstraction" (Wasmann 1905, "Ants-preface"). In contrast, Reid (1906) sees continuity between human instincts and animal instincts as involving emotional desires, but importantly plays up the enhanced capacity for learning, reasoning, and abstraction that accompanies more powerful memory and the vast store of experience ("Heredity Arg7"). Thompson also appeals to memory and learning to support his middle position between the extremes. However, as is befitting what was destined to become a long-running textbook that went into 4 editions spanning as many decades, Washburn (1908) is the most thorough, appealing to the physiologist/biologist Jacques Loeb to link behavioral variability and flexibility to "the power to learn by individual experience" ("The Animal Mind Arg13"), again revealing the importance of Loeb as a precursor of American behaviorism.

Developmental approaches

Mills (1888), Reid (1906), and Needham (1910), all explicitly advocate a developmental approach to the study of animal mind, although they do so within the framework of a strong nature-environment distinction (nowadays known as the "nature-nurture" distinction). Mills ("Nature Arg2") advocates that the task of discerning the relative contributions of nature and environment would be simpler for a dog than for a human because dogs are simpler in nature and the relevant environment for them is also simpler. Needham (1910) also argues that by varying the conditions of rearing for a litter of puppies or kittens the relative influence of environment may be determined, with consequences for theories of education ("General Arg2). Reid uses an explicit (albeit somewhat frightening) thought experiment of leaving a human infant to develop without social contact in a closet to make the point of the high dependence of humans on experience to develop mental capacities ("Heredity Arg6"), and he contrasts this with what he asserts about the ability of a tadpole to develop in isolation yet still function in nature as a typical frog. He argues that memory and reason are evolution's solution to the problem of "a natural limit to the number of instincts that may be evolved at one time in a species". Mills argues ("Nature Arg4" that the rule encoded in Lloyd Morgan's canon – to never attribute a higher psychological faculty when a lower one is sufficient – is too restrictive when it is "the truth at which we wish to get". The canon is also discussed by Washburn, who describes it as "an important counterbalance" to our most common source of error, but cautions that even when a simpler explanation suffices for a given piece of behaviour, more complex explanations "remain in the field of possibility" ("Animal Mind Arg12"). In the next argument analyzed ("Animal Mind Arg 13"), Washburn identifies "the power to learn by individual experience" as the evidence that Lloyd Morgan, Loeb, and Romanes would accept "as demonstrating the presence of mind in an animal". In the analyzed passages, Washburn does not follow the earlier authors and explicitly link learning and development. This division of learning and development appears to prefigure a significant division between comparative psychologists and development psychobiologists that would not be closed for at least fifty years (and may still not have been completely re-integrated).

Summary

These brief remarks suggest a much richer account of the content of these debates which could be developed from the materials extracted by our algorithms. And while the themes discovered should be compared with scholarly treatments of the same (e.g. Richards 1987), nevertheless we believe that despite the variations in language (vocabulary and style), the criss-crossing overlap among the arguments discovered in these books indicates that we have met a primary goal of the Digging by Debating project – that is, our methods identified pages that were thematically relevant to the stated goal of tracking the scientific and philosophical debates about anthropomorphic attributions to animals in the late 19th and early 20th centuries, thus supporting confidence in the claim that big-data analytic techniques can support close reading of texts in a content-relevant, argument-guided way.

Later we test whether similar methods applied in a more fine-grained fashion than whole pages can be used to help discover related sets of propositions belonging to a single argument.

3.4 The Argument Analysis Process: From topic modeling and mapping to automated extraction

3.4.1 The topic modeling selection process

Previously we described LDA topic modeling applied to a set of 1,315 books treating whole books as the documents, and to a smaller set of 86 books treating individual pages within them as the documents. Topic modeling can however, be applied at smaller scales, and in thinking about whether we could automate the discovery of propositions playing some sort of role within arguments, we decided to test the potential for topic modeling of single volumes treating individual sentences within a volume as the documents. Thus we ran separate LDA models for each of the six books that had been the target of our automatic page selection +

manual argument extraction process. The LDA process was the same as before with the number of topics K=20. Using topic modeling at the sentence level of analysis, it is possible to rate pairs of sentences for similarity given the probability of their being generated by the same topics.

3.4.2 The OVA modeling process

The modeling process has been similar in spirit to other, larger scale investigations into argument corpus construction such as (Reed, 2005) and (Lawrence et al., 2012). Perhaps the most significant difference in specifics lies in the level of abstraction: the discourse analysis techniques used in the argument analysis of, e.g., (Budzynska & Reed, 2012) follow the path laid down by Rhetorical Structure Theory (Mann & Thompson, 1987) in staying very close to the text, and 'reconstructing' features such as enthymematic arguments very little or not at all. In contrast, the current research has to deal with much longer spans of text and therefore necessarily departs further from it in building an abstraction of the key points. Inevitably, the challenge is to strike a balance between preserving the original text (which makes it possible to bring automated linguistic techniques to bear) and building an abstraction which allows book-sized arguments to be sketched. As the computational techniques at the fine-grain and the argument-analysis techniques at the coarse-grain improve, they are coming ever closer together.

Our approach here has been to use coarse-grained analysis as a technique for managing the challenges of real and large datasets, whilst using fine-grained analysis at a proposition-by-proposition level to train basic algorithms. With a number of straightforward mappings between the two levels (substring inclusion, for example), we can then assess the efficacy of automated techniques by comparing with the manually generated gold standard. The details of the method we have adopted and the results it has produced are detailed in §5, below. The coarse-grained analyses are available online as an AIF corpus.

3.4.3 Argument signifiers

Argument signifiers are individual or collocations of lexemes which explicitly indicate argument structure. They are referred to in the literature as discourse or argument indicators, or discourse cues or clues. The Analysis tables[15] have a column for 'Possible Signifier'. This shows a combination of general words - not necessarily at the start of the proposition - which *might* be used to indicate argument is taking place. Some propositions have no entry for 'Possible Signifier'.

Looking at the possible signifiers in the six volumes analysed, for the first volume (The Animal Mind), there are a few missing entries where it has not been possible to select some words, but there is a great variety of forms, so that 'common' signifiers appear infrequently. The richness (and verbosity) of the writing in the first volume imply much less repetition of common words and phrases. The author in the first volume is largely presenting other writers' arguments and so there is a variety of language and styles used to present the arguments. In the second volume (Psychology of Ants) there are clear argument signifiers in

[15] Not shown in this report but can be found on the link http://bit.ly/1fj38y7

about 40% of cases - such as *hence* and *therefore* - as the author makes his argument in no uncertain terms. In the third volume there are a similar number of argument signifiers, but these are less certain. It is likely, therefore, that a simple machine analysis of the selected texts, using keywords, would locate up to 40% of the propositions used. A statistical analysis of these texts, trained on similar texts, would achieve better results, but (before any machine-based results are known) we would estimate it unlikely to find more than 60% of the propositions selected in the human-based analysis of this corpus.

In general, signifiers have not been found to be reliable or ubiquitous indicators of structure, with some studies finding as few as 10% of argument structures explicitly marked with discourse material indicating structure (Moens et al., 2007). They remain, however, an important element in the arsenal of both human and machine analysts, because however rare they may be, they provide some of the clearest evidence of structure that may otherwise be rather difficult to tease out.

3.4.4 Implications for design of tools for extracting argument

If the likely success rate for machine-based extraction of arguments is less than 60% of propositions then it would not be possible to build an analyser of arguments that produced sensible results - we would need 80% or more success rate. Even if the finding of propositions were successful, we would have the problem of connecting them to make an argument structure. This would entail determining the (main) conclusion and which propositions directly support it, and which support sub-conclusions – a task that requires some level of semantics.

Finding the main conclusion is not an easy task. In the selected volumes presented here, the main conclusion is uncertain as to signifiers and position within argument, and is often difficult to determine even as a human using semantics. In our selection, the main conclusion had a clear signifier in less than 50% of the arguments, and only in the second volume was the main conclusion the last proposition of the argument with any regularity. Indeed, the more interesting and important the conclusion, the more likely it is to be situated in a complex web of propositions.

3.5 Summary of Topic Modeling, Mapping and Argument Analysis

This study has:

1. used an innovative topic modeling technique developed at Indiana to identify the volumes and then pages within those volumes which are ‘rich’ in the chosen topic;
2. used an informal discourse analysis techniques (e.g. identifying key arguments by reading through the selected pages to incrementally identify and build up argument structures;
3. used the OVA argument mapping tool to represent and map the key arguments and provide a framework for comparative analysis;
4. used a novel analysis of the mapped arguments covering the role, content and source of propositions, and the importance and meaning of arguments;

5. listed the key arguments extracted from each text using the methods above.

One potentially significant achievement in this work is that, in applying the methods above, a non-expert with limited or no domain knowledge of philosophy has identified the volumes of interest from a typical ‘Big Data Set’ (Hathi Trust) *and* identified key arguments within these texts.

4. Methodology for the Construction of an Auto-Extraction Tool

4.1 Proposition Segmentation

Proposition segmentation is carried out using a machine learning algorithm to identify boundaries, classifying each word as either the beginning or end of a proposition. Two Naive Bayes classifiers are first generated using a set of manually annotated training data. The text given is first split into words and a list of features calculated for each word. The features used are given below:

word	The word itself
length	Length of the word
before	The word before
after	The word after. Punctuation is treated as separate words so, for example, the last word in a sentence may have an after feature of '.'
pos	Part of speech as identified by the Python Natural Language Toolkit POS tagger.

Once the classifiers have been trained, these same features can then be determined for each word in the test data and each word can be classified as either 'start' or 'end'. We then run through the text and when a 'start' is reached we mark a proposition until the next 'end'.

4.2 Structure of Argument

After the individual propositions have been extracted from the text using the method described in the previous section, we focus on the structural relations between propositions. Structure is generated using the similarity of each proposition to each other proposition, determined using an LDA topic model. A topic model is first generated for the text to be studied and then each proposition identified in the test data is compared to the model giving a score for each topic. The propositions are then processed in the order in which they appear in the test data. Firstly, the distance between the proposition and it's predecessor is calculated as the Euclidean distance between the topic scores. If this is below a set threshold, the proposition is linked to its predecessor. If the threshold is exceeded, the distance is then calculated between the proposition and all the propositions that have come before, if the closest of these is then within the threshold, an edge is added. If neither of these criteria is met, the proposition is considered unrelated to anything that has gone before. In this way, the algorithm assumes that an argument is structured as a tree and presented as a depth first traversal of the tree; LDA-topic based similarity between nodes in the tree is then used to determine when to branch.

In real situations, arguments are sometimes not structured as trees (the pattern known as 'divergent argumentation,' in which a single premise supports multiple conclusions, is a good example), which is why many newer argument analysis tools such as OVA support graph-based structures. Furthermore, arguments are also not always rendered in text as depth-first traversals – with appropriate discourse cues, ('returning to the question of'; 'as discussed above'; etc.) almost any structure can be linearised into text. It is, however, true to say that most arguments are tree structured (see, e.g. Reed, 2005), and most textual renderings are depth first traversals. As a first approximation, therefore, these assumptions are reasonable ways of making the problem tractable.

5. Results of Using the Auto-Extraction Tool

As discussed in §3.1.3, the manual analysis is at a higher level of abstraction than is typically carried out in typical approaches to critical thinking and argument analysis (Walton, 2006; Walton et al., 2008), largely because such analysis is very rarely extended to arguments presented at monograph scale (see Finocchiarro, 1980, for an exception). The manual analysis still, however, represents an ideal to which automatic processing might aspire. In order to train the machine learning algorithms, however, a large dataset of marked propositions is required. To this end, the manual analysis conducted at the higher level in §2 and §3 is complemented by a more fine-grained analysis which marks *only* propositions (and not inter-proposition structure).

5.1 Comparison of the Human and Automated Analysis

5.1.1 Proposition identification

An obvious place to start, then, is to assess the performance of the proposition identification – that is, using discourse indicators and other surface features as described in §4, to what extent do spans of text automatically extracted match up to spans annotated manually described in §2 and §3? There are four different datasets upon which the algorithms were trained: (i) raw data directly from Hathi Trust – this data is contaminated with OCR errors, formatting problems and running headers, page numbers and footnotes which punctuate the text; (ii) cleaned data (with these errors manually corrected) taken only from Chapter 1; (iii) cleaned data from Chapters 1 and 2; (iv) cleaned data from Chapters 1, 2 and 4. It is important to establish a base line through processing raw (uncleaned) text, but it is expected that performance will be poor since randomly interspersed formatting artefacts (such as the title of the chapter as a running header occurring in the middle of a sentence that runs across pages) have a major impact on the surface profile of text spans used by the machine learning algorithms. All the test data is taken from Chapter 1, so it will be important to see how well multiple chapters can be used to train algorithms that might work on a small subset.

The first 'result' to note is the degree of correspondence between the fine-grained propositional analysis (which yielded, in total around 1,000 propositions) and the corresponding higher level analysis. As is to be expected, the atomic argument components in the abstract analysis typically cover more than one proposition in the less abstract analysis. In total, however, 88.5% of the propositions marked by the more detailed analysis also appear in the more abstract. That is to say, almost nine-tenths of the material marked as argumentatively relevant in the detailed analysis was also marked as argumentatively relevant in the abstract analysis. This result not only lends confidence to the claim that the two levels are indeed examining the same linguistic phenomena, but also establishes a 'gold standard' for the machine learning – given that manual analysis achieves 88.5% correspondence, and it is this analysis which provides the training data, we would not expect the automatic algorithms to be able to perform at a higher level.

Perhaps unsurprisingly, only 11.6% of the propositions automatically extracted from the raw, uncleaned text exactly match spans identified as propositions in the manual analysis. By running the processing on cleaned data (i.e. data without running headers, page numbers, footnotes and so on), this figure is improved somewhat to 20.0% for text from Chapter 1 alone (less that used for training, of course). Running the algorithms trained on additional data beyond Chapter 1 yields performance of 17.6% (for Chapters 1 and 2) and 13.9% (for 1, 2 and 4). This dropping off is quite surprising, and points to a lack of homogeneity in the book as a whole – that is, Chapters 1, 2 and 4 do not provide a strong predictive model for a small subset. This is an important observation, as it suggests the need for careful subsampling for training data. That is, establishing data sets upon which machine learning algorithms can be trained is a highly labour-intensive task. It is vital, therefore, to focus that effort where it will have the most effect. The tailing-off effect witnessed on this dataset suggests that it is more important to subsample 'horizontally' across a volume (or set of volumes), taking small extracts from each chapter, rather than subsampling 'vertically,' taking larger, more in-depth extracts from fewer places across the volume.

This first set of results is determined using strong matching criteria: that individual propositions must match exactly between automatic and manual analyses. In practice, however, artefacts of the text, including formatting and punctuation, may mean that although a proposition has indeed been identified automatically in the correct way, it is marked as a failure because it is including or excluding a punctuation mark, connective word or other non-propositional material. To allow for this, results were also calculated on the basis of a tolerance of ±3 words (i.e. space-delimited character strings). On this basis, performance with unformatted text was 17.4% – again, rather poor as is to be expected. With cleaned text, the match rate between manually and artificially marked proposition boundaries was 32.5% for Chapter 1 text alone. Again, performance drops over a larger training dataset (reinforcing the observation above regarding the need for horizontal subsampling), to 26.5% for Chapters 1 and 2, and 25.0% for Chapters 1, 2 and 4.

A further step towards generosity in comparing automatic performance with manual performance is to assess automatic proposition identification in terms of argument *relevance*. In this context, for an automatically marked proposition to be argumentatively relevant, we mean only that the text of the proposition is included somewhere in the manual analysis. The manual analysis selects only a small proportion of the text to be included in the analysis; if automatically extracted propositions are amongst that text, then the automatic extraction has at least identified material that is argumentatively relevant. The results will then stand direct comparison to the 88.5% figure, mentioned above, which represents the proportion of manually identified propositions at a fine-grained level of analysis that are present in amongst the propositions at the coarse-grained level. With unformatted text, the figure is still low at 27.3%, but with cleaned up text, results are much better: for just the text of Chapter 1, the proportion of automatically identified propositions which are included in the manual, coarse-grained analysis is 63.6%, though this drops to 44.4% and 50.0% for training datasets corresponding to Chapters 1 and 2, and to Chapters 1, 2 and 4, respectively. These figures compare favourably with the 88.5% result for human analysis: that is, automatic analysis is relatively good at identifying text spans with argumentative roles.

These results are summarised in Table 11, below. For each of the four datasets, the table lists the proportion of automatically analysed propositions that are identical to those in the (fine-grained level) manual analysis, the proportion that are within three words of the (fine-grained level) manual analysis, and the proportion that are general substrings of the (coarse-grained level) manual analysis (i.e. a measure of argument relevance).

	Unformatted	**Ch.1**	**Chs.1 & 2**	**Chs.1, 2 & 4**
Identical	11.6%	20.0%	17.6%	13.9%
±3 words	17.4%	32.5%	26.5%	25.0%
Substring	27.3%	63.6%	44.4%	50.0%

Table 11. *Results of automatic proposition processing*

5.1.2 Argument structure identification

Clearly, identifying the atoms from which argument ‘molecules’ are constructed is only part of the problem: it is also important to recognise the structural relations. Equally clearly, the results described in §5.1.1 have plenty of room for improvement in future work. They are, however, strong enough to support further investigation of automatic recognition of structural features (i.e., specifically, features relating to argument structure).

In order to tease out both false positives and false negatives, our analysis here separates precision and recall. Furthermore, all results are given with respect to the coarse-grained analysis of §2 and §3, as no manual structure identification was performed on the fine-grained analysis.

As described in §4, the automated structure identification currently returns connectedness, not direction (that is, it indicates two argument atoms that are related together in an argument structure, but do not indicate which is premise and which conclusion). The system uses propositional boundaries as input, so can run equally on manually segmented propositions (those used as training data in §5.1.1) or automatically segmented propositions (the results for which were described in Table 5). In the results which follow, we compare performance between manually annotated and automatically extracted propositions. Figures 4 and 5 show sample extracts from the automated structure recognition algorithms running on manually segmented and automatically segmented propositions respectively.

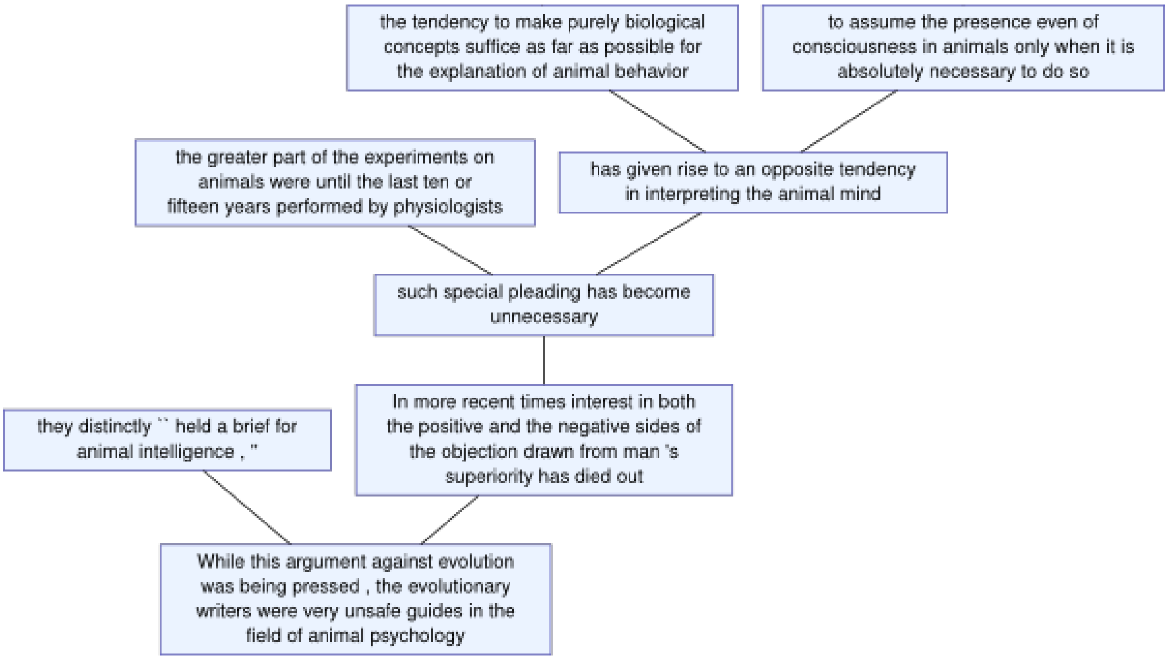

the tendency to make purely biological concepts suffice as far as possible for the explanation of animal behavior
to assume the presence even of consciousness in animals only when it is absolutely necessary to do so
the greater part of the experiments on animals were until the last ten or fifteen years performed by physiologists
has given rise to an opposite tendency in interpreting the animal mind
such special pleading has become unnecessary
they distinctly `` held a brief for animal intelligence , ''
In more recent times interest in both the positive and the negative sides of the objection drawn from man 's superiority has died out
While this argument against evolution was being pressed , the evolutionary writers were very unsafe guides in the field of animal psychology

***Figure 4**. Example of automated structure recognition using manually identified propositions*

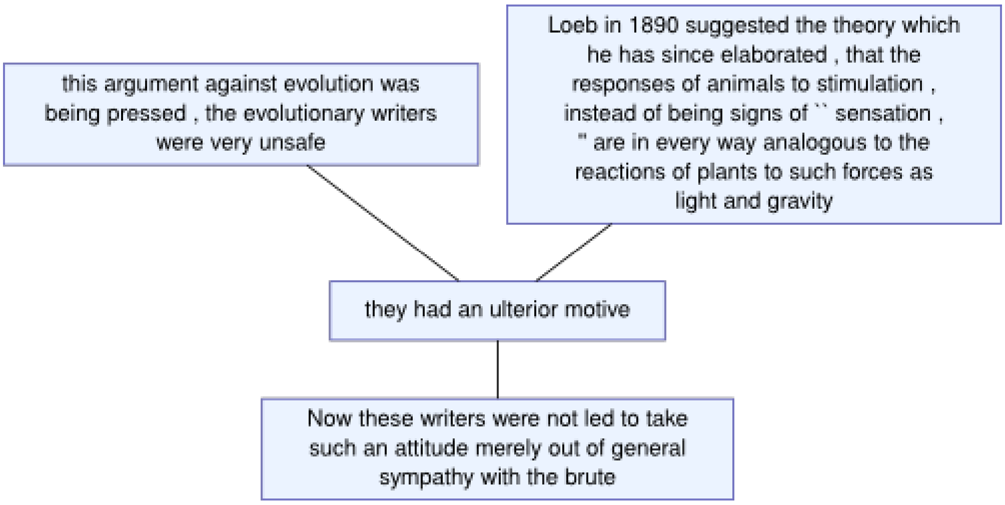

this argument against evolution was being pressed , the evolutionary writers were very unsafe
Loeb in 1890 suggested the theory which he has since elaborated , that the responses of animals to stimulation , instead of being signs of `` sensation , '' are in every way analogous to the reactions of plants to such forces as light and gravity
they had an ulterior motive
Now these writers were not led to take such an attitude merely out of general sympathy with the brute

***Figure 5.** Example of automated structure recognition using automatically identified propositions*

For all those pairs of (manually or automatically) analysed propositions which the automated structure recognition algorithms class as being connected, we examine in the manual structural analysis connectedness between propositions in which the text of the analysed propositions appears. Thus, for example, if our analysed propositions are the strings *xxx* and *yyy*, and the automated structure recognition system classes them as connected, we first identify the two propositions (P1 and P2) in the manual analysis which include amongst the text with which they are associated the strings *xxx* and *yyy*. Then we check to see if P1 and P2 are (immediately) structurally related. For automatically segmented propositions, precision is 66.6% and recall 100.0%, whilst for manually segmented propositions, precision is 33.3% and recall 18.2%. For automatically extracted propositions, the overlap with the coarse-grained analysis was small – just four propositions – so the results should be treated with some caution. Precision and recall for the manually extracted propositions however is based on a larger dataset (*n*=26), so the results are disappointing. One reason is that with the manual analysis at a significantly more coarse-grained level, propositions that were identified as being structurally connected were quite often in the *same* atomic unit in the manual analysis, thus being rejected as a false positive by the analysis engine. As a result, a more liberal analysis was conducted, in which success is recorded if either

(a) for any two manually or automatically analysed propositions (p1, p2) that the automated structure recognition indicates as connected, there is a structural connection between manually analysed propositions (P1, P2) where p1 is included in P1 and p2 included in P2

or

(b) for any two manually or automatically analysed propositions (p1, p2) that the automated structure recognition indicates as connected, there is a single manually analysed proposition (P1) where p1 and p2 are both included in P1

Under this rubric, automated structure recognition with automatically segmented propositions has precision of 66.6% and recall of 100% (but again, only on a dataset of *n*=4), and more significantly, automated structure recognition with manually segmented propositions has precision 72.2% and recall 76.5%. These results are summarised in Table 12. The results are encouraging, but larger scale analysis is required to further test the reliability of the extant algorithms.

	Automatically segmented propositions	**Manually segmented propositions**
In separate propositions	*n*=4, P=33.3%, R=50.0%	*n*=26, P=33.3%, R=18.2%
In separate or the same proposition	*n*=4, P=66.6%, R=100.0%	*n*=26, P=72.2%, R=76.5%

Table 12. *Results for analysis of propositions*

5.2 Automated Analysis: Implications for further work

With fewer than one hundred atomic argument components analysed at the coarse-grained level, and barely 1,000 propositions at the fine-grained level, the availability of training data is a major hurdle. Developing these training sets is demanding and extremely labour intensive. One possibility is to increasingly make available and re-use datasets between projects. Infrastructure efforts such as aifdb.org make this more realistic, with around 15,000 analysed propositions in around 1200 arguments, though as scale increases, quality management (e.g. over crowdsourced contributions) becomes an increasing challenge.

With sustained scholarly input, however, in conjunction with cross-project import and export, we would expect these datasets to increase 10 to 100-fold over the next year or two, which will support rapid expansion in training and test data sets for the next generation of argument mining algorithms.

6. Applications to, and Technical Development of PhilPapers

6.1 Application of Topic Modeling Tools to PhilPapers

While using the interface of science and philosophy in the area of animal psychology as a test case, the project also aimed to generalize the use of modeling tools and automatic content extraction techniques to other fields and subjects and to make these tools available for other projects to use. We made progress on this front by adapting the tools and applying them to PhilPapers, the primary citation index and open access archive in philosophy[16].

David Bourget and his team at PhilPapers adapted and used the topic modeling tools developed by Allen's team to model the PhilPapers corpus. A wide range of model types were used, with numbers of topics varying from 30 to 1200. SHARCNET (the Ontario supercomputing network) was used to perform the largest modeling experiments.

The models generated were to put to the test in two tasks: feature reduction for categorization, and similarity matching. In the first task, an item's topics replace its text for the purposes of automatic categorization into one of PhilPapers' 5000 or so categories. PhilPapers contains approximately 1,000,000 published items (books or articles), about half of which are already categorized in the 5,000 categories. The already-categorized items provided the training set for our experiments.

Bayesian, Support Vector Machines (SVM), and combined Bayesian-SVM classifiers were tried with a range of parameters. The performance of these classifiers was compared to that of control classifiers that were otherwise equivalent but used text instead of topics. Topics were found to slightly decrease precision but slightly increase recall with the Bayesian classifier. The performance was comparable in the best case scenarios. The main benefit was a decrease in training time from multiple weeks to a few days (due to the use of a much small number of features), which will make it possible to continue to use the whole PhilPapers corpus as training set while it grows. Precision and recall were significantly decreased with SVM-based classifiers, making topic modeling a poor feature reduction technique for these classifiers.

We also obtained mixed results with the second task, similarity matching. In this task, a program is given an article from PhilPapers and must find the ten articles in the corpus that are most similar to it. We tested over twenty different configurations of models and selection techniques using these models. The general approach was to return the articles that share the most topics with the supplied articles, but different techniques were used in an attempt to improve the results. For each approach tested, we generated 100 sample selections for

[16] http://philpapers.org

random articles from the corpus. These selections were given to two philosophy graduate students at the University of Western Ontario to evaluate separately. Twelve experiments of this type and many smaller experiments were conducted. This process enabled us to fine-tune the system to the point where students report that its performance is equal to or better than that of comparable recommendation systems on Amazon and Google Books. The best similarity matching technique combines topic-based estimates of similarity with literal keyword matching.

Both the similar article recommendation system and the classification system based on topic modeling are currently being moved to production on PhilPapers. Similar articles can be seen on PhilPapers record pages[17].

6.2 Application of Mapping Techniques to PhilPapers

The UCSD map of science technology was also applied to the PhilPapers corpus. Final results are pending and will be announced on the project's site.

6.3 Related Technical Developments at PhilPapers

In order for the current application to yield the best results on PhilPapers, it was necessary to fill certain gaps in its corpus that were due to biases in the construction of the index. A major data gathering exercise was launched. We developed a new set of data trawling tools which enabled us to double the size of PhilPapers index, from approximately 500,000 items at the beginning of the project to approximately 1,000,000 items at the time of writing.

6.4 PhilPapers Software

All the PhilPapers-related software developed as part of this project is available under the GNU Public License via GitHub as part of the xPapers application framework. The topic modeling tools (including categorizations and similarity matchers, with all topic modling parameters configured for optimal performance) are currently part of the "LDA" branch of xPapers: https://github.com/xPapers/xPapers/tree/lda

The crawling tools are part of the "master" branch of xPapers:
https://github.com/xPapers/xPapers/tree/master

[17] Such as this one: http://philpapers.org/rec/BOUCIU

7. Discussion

The automated content selection and categorization work described in Sections 2.2 and 6 demonstrated the feasibility and reliability of large-scale, fined-grained topic-based categorization across a range of topics in science and philosophy using documents defined at a variety of scales (whole books, book pages, and individual sentences in books from the Hathi Trust collection, article abstracts on PhilPapers). Categorization and selection are essential first-steps in the Digging by Debating process that this project aimed to prototype for large data sets. The project demonstrated that these tasks can be successfully completed. The techniques demonstrated promise to enable a new level of organization of online collections along thematic lines. This technology is already generating vastly improved categorization on PhilPapers.

The Argument study described in Section 2.3 and 3 provided a number of key insights about the nature of argument in the texts we studied, the 'rendering' of the arguments within these historical texts and the challenges for automated argument identification, extraction and representation. It is also an innovative approach to digging into data, as the emphasis here was to 'drill down' into the argumentative meaning and understand it better, before tackling the problem of how we can dig into it with the assistance of algorithmic techniques. Or, putting this another way, we aimed to carefully characterise and understand the nature of 'argumentative text as data' before digging into it. This can be seen as part of a process of problematising the design space, to better understand the problem before proposing technical solutions. A number of insights emerged from performing this human analysis of texts that were 'delivered' by our topic modeling techniques and then mapped in argumentative terms through the formal modeling tool - OVA.

Firstly, the LDA topic modeling technique that was used was clearly successful in identifying the texts (chapters and pages) that contained the 'stuff' of arguments linked to the keywords and topics that were searched for. These could be sorted through rankings that allowed just the 'topic rich' texts to be the focus of further analysis. This is very valuable in itself, as it allowed us to extract the most argument rich texts from a 'big data' text repository (Hathi Trust). Secondly the (human) argument identification and analysis produced 60 argument maps (in OVA) that served as useful representations in themselves and clearly defined objects for further analysis and comparison. Also most interesting was the insight gleaned from the process of manually identifying, interpreting and mapping the arguments, if this process is conceived as similar to one of knowledge engineering. The argument forms and structures could not be immediately identified and extracted by a human reader, but the process followed an 'algorithm' of:

1. read through the text to get a broad brush nature and meaning of the arguments at play;
2. find the linguistic argumentative markers to further define and represent the arguments;
3. once an argument is constructed - hypothesis test backwards - to see if the apparent argument is rendered in the texts;
4. refine or re-construct the argumentative text;

5. iterate through 1- 4 as required until the map seems stable.

Conducting this process showed that, in the texts we studied, arguments were rarely clearly structured and defined explicitly in the texts. Instead arguments seem to ‘come alive’ through the practice of interpreting, understanding and (re)constructing them. This raises the somewhat controversial and polemical question: ‘Do arguments actually exist in clearly defined forms within (certain) texts? Or, do arguments only take form when we focus on understanding them?’ In more precise terms: Do arguments only exist within the interaction between mind and text? While these questions are too big to be answered by our study and project here and now, their potential validity as questions are, we argue, supported. And these questions clearly link to the challenges we experienced with defining argument structure and producing tools that could automatically identify, construct and map them.

Argument structure is notoriously difficult for people, even after training, to determine. Values for Cohen’s kappa statistic are typically very low, demonstrating poor inter-coder reliability and thence poor reproducibility. Of course, this should come as no surprise when even textbooks of argument analysis disagree with one another on the simplest of examples. Yet everyone agrees that there is some deep structure which is recoverable, and that recovering it or reconstructing argument structure is an important scholarly activity that aids comprehension of difficult texts, and that the skill of doing so is a legitimate aspect of mental agility, essential for informed citizenship and deserving of inclusion in university syllabi. For some logicians, linguists and philosophers, the focus is on extensive reconstruction of important examples (e.g. Zeno’s argument against the possibility of motion, Anselm’s ontological argument for the existence of God, or Descartes’ argument against radical skepticism), filling in ever more material to render arguments, in the limit, as deductively valid (but which they almost never are in their raw forms). For others, argument analysis is a much more fluid and loose phenomenon in which it is acceptable for there to be multiple defensible analyses of a given text. Despite the postmodernist hue to this end of the spectrum, it is consistent with more hard-nosed research in pragmatics in particular, where techniques such as Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) defends the use of ‘plausibility judgements’ in determining reasonable RST analyses (but RST also suffers from poor kappa scores). Large-scale enterprises that aim at supporting analysis of substantial datasets (including aifdb.org, debategraph.org and argunet.org) typically push very much towards the latter end of the scale as a matter of necessity. In general, progress towards more effective use of massive text repositories by humanities scholars will require a combination of computational techniques, digital curation by experts, and social computing, linking dialogue to semantic searching and extraction (e.g. Ravenscroft, Braun and Nelker, 2011).

Despite the challenges however, this project has demonstrated that the situation is far from hopeless for computational techniques. First, propositional analysis is relatively straightforward and reasonable performance can be achieved with modest feature sets, low complexity learning, and well established techniques. This outcome is to be anticipated, given that kappa values for propositional boundary marking is typically quite high. For structural analysis, we have only made modest inroads on the basis of assumptions of linguistically rendered depth-first traversal of tree-structured arguments, but even so, the technique is successfully identifying some of the argumentative relations in the text. This establishes a baseline for developing a battery of techniques which can be combined to build more robust techniques for argument mining. The key bottleneck is analysed data. The two leading data sets are aifdb.org at Dundee and debategraph.org, both of which have of the order of 10^5 analysed arguments (for comparison, the Digging By Debating project has added around 350). Of the two, only aifdb.org is fully open and accessible for academic research. But it cannot support the development of computational learning techniques for the wide variety of argument structures without growing significantly – to at least 10^7. Tying in to crowdsourced, scalable analysis techniques is going to be key in driving towards this practical goal and unlocking further statistical techniques for structure extraction.

Further development and understanding of the LDA methods and their applications can also be anticipated. The methods described above selected the number of topics to use for the modeling in an ad hoc fashion. Although there exist estimates of optimal numbers of topics available within computer science (Heinrich 2009; Arun et al. 2011), these estimates are based on measures that don't take into account the kinds of applications that we have been pursuing. Most of these methods look for an inflection point of diminishing discriminatory returns, such that, e.g., 20 topics might capture 85% of the variance among the texts, whereas it might take another 60 topics to capture the next 10%. However, although 20 topics might be adequate for a high-level classification scheme, we believe that a larger set topics can enable finer-grained selection at multiple scales – a set of 1,315 books modeled at the volume level, a set of 86 books modeled at the page level, and single books modeled at the sentence level – and that this may be especially true for the discriminations necessary at the sentence/argument level. We are currently poised to make more systematic comparisons of the interaction between number of topics and scale of analysis, but to do so was outside the scope of the current Digging by Debating project.

8. Conclusions and Recommendations

8.1 Conclusions and Recommendations from the Human Study

Use of the Topic Modeling methodology appears to have been quite successful in selecting volumes, though there was some concern that the number of volumes available that meet the criteria (i.e. 20-40 pages) of on-topic material is low. Key to the selection is the choice of abstracted topic - the selection could vary dramatically if a different choice was made. The analysis scheme devised appears to discriminate quite well between different writing / argument styles.

The following points can be made:

1. The coding scheme (proposition content, source and role) appears to discriminate quite well between volumes, as it highlights significant differences, mainly between the first three volumes analysed, and these differences align with our impressions as readers of the volumes.
2. The number of arguments that can be mapped from volumes varies considerably – in this study from two to seventeen – even with the 20-40 on-topic pages criteria.
3. A particular author's style seems to be apparent within about five arguments analysed.
4. Rather than mapping a set number of pages, the focus should be in extracting a small number of ‘important’ arguments from the volume.
5. More work on recognising signifiers may be useful. Note that the force of signifiers used may not imply ‘sound’ arguments. Some of the more subtle and complex arguments lack the strong signification found in simpler arguments backed by strongly held opinion (compare the first and second volumes).
6. Searching for conclusions in the texts and working backwards (and forwards) to extract argument is probably the most efficient method for a human reader. Usually, conclusions are signposted most clearly and the human reader must then make sense of the surrounding text to determine what is relevant to the conclusion. A similar strategy might be adapted to work for automatic extraction.

8.2 Conclusions and Recommendations on Auto-Extraction Tools

Key conclusions drawn:

1. Infrastructure and tooling is maturing well to allow networks of researchers to collaborate on argument analysis at scale.
2. Automatic segmentation of propositions in a text on the basis relatively simple features at the surface and syntactic levels is feasible, though generalisation between chapters, volumes and, ultimately, genres, is extremely demanding.
3. Automatic identification of at least some structural features of argument is surprisingly robust, even at this early stage (performance and recall both over 70%), though more

sophisticated structure such as inferential directionality and inferential type is likely to be much more challenging.
4. It is feasible to connect automatic segmentation and automatic structure recognition, though much more data is required to test its applicability at scale.

Key recommendations for future work:

1. Significantly expanded datasets are crucial which will require networks of researchers to share the load of human analysis.
2. Propositional segmentation could be improved by making more thorough use of syntactic information such as clausal completeness (rather than just POS tags).
3. With a significant foundation for argument structure analysis laid by this project, future work can focus on extending and refining sets of algorithms and heuristics based on both statistical and deep learning mechanisms for exploiting not just topical information, but also the logical, semantic, inferential and dialogical structures latent in argumentative text.

8.3 Conclusions and Recommendations on Topic Modeling Tools

Key conclusions:

1. LDA-based topic modeling is a better means of identifying works relevant to a topic than TF-IDF and BEAGLE modeling.
2. LDA-based topic modeling can enable a researcher to narrow down reading material relevant to a topic by more than 99%.
3. Topic modeling can be an effective means of reducing training time for Bayesian classifiers without significantly affecting precision or recall, whereas precision and recall is significantly degraded using SVM-based classifiers.
4. Topic modeling can be an effective method for matching articles by similarity. Optimal results are obtained in combination with keyword matching.

Key recommendations for future work:

1. Future work could further advance the current work by studying the effectiveness of topic modeling technology as a tool for generating research insights in the humanities.
2. An open platform enabling researchers without technical expertise to analyze corpora using LDA could open new research directions.

8.4 General Conclusions

Our digging by debating project has tackled a very ambitious and complex problem, of linking macro visual views of science-philosophy data and state-of-the-art topic modeling and searching to semantically rich analysis and processing (based on argument) of the data. Our project has made significant steps forward in these areas and their interconnection, and produced a constellation of loosely integrated tools in this respect.

Our project is showing for the first time how big data text processing techniques can be combined with deep structural analysis to provide researchers and students with navigation and interaction tools for engaging with the large and rich resources provided by datasets such as the Hathi Trust and PhilPapers. Ultimately our efforts show how the computational humanities can bridge the gulf between the “big data” perspective of first-generation digital humanities and the close readings of text that are the “bread and butter” of more traditional scholarship in the humanities.

9. References

1. R. Arun, V. Suresh, C. E. Veni Madhavan, M. N. Narasimha Murthy (2010). “On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations”, *Advances in Knowledge Discovery and Data Mining – Lecture Notes in Computer Science* Volume 6118, 2010, 391-402. (online at http://link.springer.com/chapter/10.1007%2F978-3-642-13657-3_43).

2. Katy Börner, Richard Klavans, Michael Patek, Angela Zoss, Joseph R. Biberstine, Robert Light, Vincent Larivière and Kevin W. Boyack (2012). "Design and Update of a Classification System: The UCSD Map of Science", *PLoS One* 7 (7).

3. Chesnevar, C., McGinnis, J., Modgil, S., Rahwan, I., Reed, C., Simari, G., South, M., Vreeswijk, G. and Willmott, S. (2006). "Towards an Argument Interchange Format", *Knowledge Engineering Review*, 21 (4), 293-316.

4. Digging by Debating Abstract (online at http://diggingbydebating.org/wp-content/uploads/2013/03/DiggingByDebatingAbstract.pdf).

5. Maurice A. Finocchiaro (1980). *Galileo and the Art of Reasoning: Rhetorical Foundations of Logic and Scientific Method* (Boston Studies in Philosophy of Science, vol. 61.). Dordrecht: Reidel (now Springer).

6. Heinrich, G. (2009). “A generic approach to topic models”, *Proceedings ECML/PKDD, September 2009* (online at http://www.arbylon.net/publications/mixnet-gibbs.pdf).

7. Lawrence, J., Bex, F., Reed, C. & Snaith, M. (2012). "AIFdb: Infrastructure for the Argument Web" in *Proceedings of the 4th International Conference on Computational Models of Argument* (COMMA 2012), IOS Press, Vienna, 515-6.

8. Mann, William C. and Sandra A. Thompson (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization.", *Text 8* (3): 243-281.

9. Stephen Merity, Tara Murphy and James R Curran (2009). “Accurate Argumentative Zoning with Maximum Entropy models”, *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, ACL-IJCNLP 2009, 19-26.

10. Marie-Francine Moens, E Boiy, Raquel Palau and Chris Reed (2007). “Automatic detection of arguments in legal texts” in *Proceedings of the 11th international conference on Artificial intelligence and law*, 225-230. ACM.

11. Jaimie Murdock, Robert Light, Colin Allen and Katy Börner (2013). "Mapping the Intersection of Science & Philosophy", *Joint Conference on Digital Libraries (JCDL)* (online at http://cns.iu.edu/docs/publications/2013-murdock-digging-jcdl.pdf).

12. David Newman, Arthur Asuncion, Padhraic Smyth, Max Welling (2009). “Distributed Algorithms for Topic Models”, *Journal of Machine Learning Research* (10) 1801-1828.

13. Raquel Mochales Palau and Marie-Francine Moens (2011). “Argumentation mining”, *Artificial Intelligence and Law*, 19 (1), 1-22.

14. Pilkington R.M. (1999). *Analysing Educational Discourse: The DISCOUNT Scheme, Version 3,*

January 1999. CBL Technical Report. no. 99/2.

15. Ravenscroft, A. & Pilkington, R.M. (2000). “Investigation by Design: Developing Dialogue Models to Support Reasoning and Conceptual Change”, *International Journal of Artificial Intelligence in Education: Special Issue on Analysing Educational Dialogue Interaction: From Analysis to Models that Support Learning,* Vol. 11, Part 1, 273-298, ISSN 1560-4292.

16. Ravenscroft, A. (2007). “Promoting Thinking and Conceptual Change with Digital Dialogue Games”, *Journal of Computer Assisted Learning (JCAL)*, Vol. 23, No 6, 453-465.

17. Ravenscroft, A. Wells, S., Sagar, M. & Reed, C. (2009). “Mapping persuasive dialogue games into argumentation structures”, *Symposium in Persuasive Technology and Digital Behaviours Intervention, Artificial Intelligence and Simulation of Behaviour (AISB) Convention 2009,* Edinburgh, April 6-7, Scotland, UK

18. Ravenscroft, A., Braun, S. & Nelker, T. (2010). “Combining Dialogue and Semantics for Learning and Knowledge Maturing: Developing Collaborative Understanding in the ‘Web 2.0 Workplace’”, in *Proceedings of International Conference on Advanced Learning Technologies (ICALT) 2010*, July 5-7, 2010, Sousse, Tunisia.

19. Ravenscroft, A. (2011). “Dialogue and Connectivism: A new approach to understanding and promoting dialogue-rich networked learning”, *International Review of Open and Distance Learning, Special Edition: Connectivism: Design and delivery of social networked learning,* (eds) George Siemens and Gráinne Conole Vol 12, No. 3: 139-160.

20. Richards, R. (1987). *Darwin and the Emergence of Evolutionary Theories of Mind and Behavior.* Chicago: University of Chicago Press.

21. Walton, D. (2006). *Fundamentals of Critical Argumentation*. Cambridge: Cambridge University Press.

22. Walton, D., Reed, C. and Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press.

Section B: Practical Management and Issues

1. How the project progressed over time, and how we managed it

The project contained a number of management challenges.

- How to perform a joint project with independent 'halves' that have different funding and project management models?
- How to hold regular working meetings to maintain national and international co-operation and joint working with limited cross-over in typical working hours?
- How to procure a consortium agreement that covers a joint project but is only 'legally binding' on one half?
- How to articulate data access issues that are evolving as the project progresses?
- How to specify the joint working in a way that captures joint responsibility and also 'legal' independence?
- How to dig into data that is freely available to one partner (the US) but restricted to the other (UK)?

The distributed and yet complex nature of the project meant that it took some time to develop effective working and meeting practices. Also, during 2012 a lot of effort was directed to solving the management and contractual challenges (that was particularly complicated in connection with data issues), project problem identification and adaptation of the existing tools. A literature review was produced by the Dundee team and a cognitive science style walk-through of the argument extraction problem by the UEL team, using the Indiana inPho tools. From October 2012 we initiated weekly (Skype) meetings between UK based team members, with full team meetings every two weeks.

A number of administrative problems at the beginning seemed to soak up an undue amount of effort. One was copyright differences between using the chosen corpus in the UK and USA, a much restricted content was available in the UK. Another was the whole Consortium agreement, with additional complications due to David Bourget's leaving the University of London – negotiating an agreement between the two parties took twelve months. Our solution to the copyright problems was access to the data through Indiana's servers where it was needed by the UK-based team in the short-term, while negotiating for equivalent access directly with the body concerned (HTRC) for the longer term.

We examined the tools on hand to the team during autumn 2012 with a view to adaptation. In particular, we spent time looking at the Topic Modeling tools and using iPython Notebooks to create and record sessions. Because of the potential computing load on some tasks (such as model training) running into hours, even days, on Indiana's servers, runs had to be carefully coordinated with Indiana's own requirements. During this time the Indiana team made numerous improvements and specialisations to the Topic Modeling tools.

At this stage there was much discussion about the sort of data we were dealing with and type and scope of problems that would arise from it. One of the problems was in consistent markup of argument. Two philosophers in the DbyD team took on independent markup of argument in two short (modern) texts. Sufficient variation occurred (about a third of markup did not overlap) that we could see that this alone would be a problem (i.e. deciding what the ‘target’ actually is).

So, in late 2012, we started a manual study of argumentative passages in a small number of texts to ‘get to grips with real data’. We decided upon our topic and honed in on a likely set of (1315) volumes to search using Topic Modeling by the end of 2012. A selection process using Topic Models was completed in February 2013 and the UEL team began markup of three volumes (later six) using OVA+. The result was an Interim Report delivered in June 2013 to the team, and which has been incorporated into this report. Once this was delivered, the plan for how the project should be finished off was mapped out at a face-to-face team meeting hosted by Dundee.

2. Evidence for how the project has improved the research environment

At Indiana University

Vector space modeling software written at Indiana has been published under an Open Source license at http://github.com/inpho/vsm and iPython notebook examples at http://github.com/inpho/vsm-demo-notebooks, deployed at Indiana, Dundee, and Western Ontario. Allen’s InPhO group has also established a working relationship with Stephen Weldon’s IsisCB history of science bibliographic project at the University of Oklahoma, where the software will also be deployed.

At University of Western Ontario

The PhilPapers index, a central component of the research environment in philosophy, has been greatly improved in content and organization. The index has doubled in size thanks to the work of Bourget and his team as part of this project. The index also now offers LDA-backed “similar papers” links, and an improved categorization workflow is enabling PhilPapers’ 500-strong team of editors to quickly populate PhilPapers’ structured bibliography of the field, a key resources for all philosophers.

At University of Dundee

Automatic extraction of meaningful information from natural text remains a major challenge facing computer science and AI. The Dundee approach to argument mining has deployed tools and techniques in novel ways, in particular to link propositions within argument structure. Refinement of techniques and integration of composite tools is required, along with the necessary volume of training data to better distinguish propositional entities, in order to make a serious assault on the problem. We have, however, laid down some essential ground-works for future research in argument mining.

At University of East London (UEL)

The work in this project on analysing argumentation and mapping argument using formal tools builds upon a large body of previous work into Digital Dialogue Games for radical Technology Enhanced Learning (TEL), that is now a central strand of research within the International Centre for Public Pedagogy (ICPUP, www.uel.ac.uk/icpup/) at UEL.This is a Design Based Research approach to understanding and analysing argumentation that has led to the development of a robust and usable tool, called InterLoc (see www.interloc.org.uk). This is an application that instantiates dialogue games that can be designed for various purposes, usually to promote critical or creative discussion, reasoned dialogue and collective inquiry (e.g. amongst students). InterLoc has recently been integrated into UEL's VLE, and continues to be used internationally, to support critical discussion and reasoning online, for numerous learning applications.

So, this project has considerably enhanced the capacity for research into argumentation within ICPuP and UEL, through placing it clearly within the Big Data, Digital Humanities and computational argument discourses. It has also broadened ICPuP's remit and skill set, through the development of a novel technique for identifying and mapping arguments within historical texts that are systematically selected from massive document respositories.

3. Meetings and important milestones

London meeting January 2012

Amsterdam meeting June 2012

Milestone October 2012 - First deliverable to Jisc, consisting of literature review and cognitive science style walkthrough of the problem area, plus workplan.

Milestone February 2013 - Topic modeling process completed, a selection of pages sent to UEL by Indiana.

Meeting June 2013 - at Dundee, manual study interim report delivered by UEL, mapping of project to completion agreed.

Meeting October 2013 - Montreal Digging into Data meeting, presentation of results so far.

Milestone January 2014 - project end.

4. Lessons learned from the project

In such a project as this, which was ambitious and had complexities, it inevitably takes time to achieve effective collaborative working. At first, the internationally distributed nature of the team mitigated against in-depth involvement (for sufficient time) of the team members for ideas to be evolved and collaboratively developed. There are times when face-to-face meetings are not easily substituted, and with hindsight the project may have picked up more quickly and more progress made in developing a ‘route’ to completion if there had been more face-to-face meetings early on. We would distinguish between ‘business’ meetings, with a full agenda and no surprises, and the ‘brainstorming’ meeting which only has one or two ideas on the agenda. The former is easily achievable with an online audio session (as we had), but the latter is not, for various reasons to do with being human. We would therefore plan immersive, creative sessions between the teams early on, probably in face-to-face sessions (which would require a larger travel budget).

The size and variety of the data set caused a lot of speculation by the team – an unknown that remained unknown for too long. However, a) we were wary of ‘jumping in with both feet’ analysing data before we knew what the data was, and b) we needed a good procedure for selecting data. The data selection process took time to refine, and we did not want to waste time to have to redo analysis if the selection procedure was found wanting. Once we had a good selection procedure we started the manual analysis of volumes and learned what we might have to do using tools for auto-extraction. As we were expecting, we were able to make real progress once we had real data to refer to and discuss. However, new ways of using the modeling tools occurred to us right up to June 2013, for example, the proposition similarity technique. There is a tension between casting a wide and thorough search, which requires tool development and innovation, and ‘getting on with it’ using the tools available.

5. Software (including IP arrangements), algorithms, and techniques developed, and how these might be sustained over time

Visualisation software will continue to be maintained and developed by Katy Börner at Indiana. https://inpho.cogs.indiana.edu/scimap/jisc/ is the link for this project, but also check with website for updates.

Topic modeling software will continue to be maintained and developed by Colin Allen at Indiana. Open Source software can be found at http://github.com/inpho/vsm/

Argument extraction tools and algorithms - Dundee Team. Open Source repository can be found at https://github.com/argdundee/DbyD.

Major additions to the xPapers platform used by PhilPapers were made and will be maintained by David Bourget and his team at Western Ontario as part of their mandate to maintain PhilPapers. Open Source software (GNU Public License) can be found at https://github.com/xPapers/xPapers/

InterLoc Digital Dialogue Game tool, http://www.interloc.org.uk/, Open Source software at sourceforge/interloc.

For all links please check with the diggingbydebating.org website which has the latest version.