Notes on the Domesday Dataset Paper

Joseph Reddington

February 25, 2013

The following paper is designed to introduce the Domesday Dataset and make it available to other researchers. The paper's intention is not to give a precise estimate of spend, although we expect that other researchers will be able to use the data to do so. The paper includes include, *as an example* a naive estimate of how such a spend could be calculated, but I would rather that the example calculations were only quoted, if at all, with the range of caveats that the paper includes and in the spirt that the paper gave them.

More to the point, I intend to update this pre-proofread version of the paper as more data comes in and more items of equipment are priced and accurately intended, so the example numbers are likely to change in any case - quoting them without caveat might be unwise Joe

THE DOMESDAY DATASET: linked open data in disability studies

Journal name
000(00):1–13

DOI:doi number
http://mms.sagepub.com

Abstract

Augmentative and alternative communication (AAC) devices provide the ability for many people with disabilities to make themselves understood. For the large proportion of users with an intellectual disability, these devices may be their only means of communication. Estimates of the number of AAC devices in use are vague and lack transparency. This prevents researchers from answering key questions like "How many people per 1,000 are using electronic AAC?", and "On average, how much money would fund a person's AAC use for 12 months?".

This work presents the 'Domesday Dataset', allowing researchers to answer research, policy, and prevalence questions on the landscape of AAC use within the UK. This dataset was constructed by making several hundred Freedom of Information (FOI) requests. These requests asked public bodies who purchase AAC devices in the UK to supply details of every AAC purchase since 2006 including make, model, and year of purchase.

Keywords

AAC, intellectual disability, Freedom of Information, transparency in data.

1. Introduction

Augmentative and alternative communication (AAC) devices provide the ability for many people with disabilities to make themselves understood. For the large proportion of AAC users with an intellectual disability, it may well be the only way. Estimates of both user provision and the number of AAC devices in use are vague and lack transparency. This lack of data prevents researchers from answering key questions in the field and causes difficulties at the research, governance and political levels.

Speech-language therapy in general and AAC in particular have an active and vibrant research area, but the return for unit cost is restricted by the inability to show how applicable individual research efforts are to the wider community: the two . Consequently, it is difficult for researchers and funding bodies to judge the potential effects of individual research strands, which could result in the oversight of many innovations of clear immediate benefit because the user-domain was incorrectly thought to be too small to support the change.

At the governance level, Speech and Language Therapists (SLTs) are required to advise for each individual user which AAC device would best fit their requirements with no definite insight into how SLTs in other areas are advising their users. At the political level, there has been some lobbying by campaigners in the UK to see communication as a right within the remit of the National Health Service (NHS). Similarly, recent healthcare legislation in the US has raised the possibility that AAC devices may be part of a universal health provision. However, campaigners have no way of answering the question "How much will this cost?" with any hard evidence.

A particularly frustrating aspect of this lack of data is that (prior to the explosion in AAC apps following the launch of the Apple iPad) there are relatively few manufacturers of AAC devices, all of which are quite capable of releasing this data, if not to the general public, then to a limited number of researchers. Unfortunately, most manufacturers are, understandably, unwilling to release sales figures on their products due to issues of competition.

Although a global census of AAC devices is a challenging goal, it may be possible to provide datasets for individual countries. This paper describes the construction of such a dataset for the United Kingdom. This is possible because the majority of AAC devices in the UK are supplied by a set of public bodies, which purchase directly from the manufacturers (See Section 2). Our work constructed a dataset by making several hundred Freedom of Information (FOI) requests, which asked public bodies to supply lists of every AAC device purchased since 2006 including make, model, and year of purchase. The resulting repository of information on AAC provision within the UK is augmented with both the raw data and original correspondence with the each public body to form the 'Domesday Dataset', which is named after the great English survey of 1086. Table 1 shows a sample of the resulting data.

The integrity and reliability of the original construction of the dataset has been reinforced by public scrutiny of each version of the data. Following the open data agenda of Berners-Lee (2010), members of the public, AAC specialists, researchers and manufacturers were encouraged to investigate the dataset, and eliminate inconsistencies and errors in the data, resulting in an accurate and wide-ranging dataset. This dataset is not only of use for shaping UK policy and research, but as a snapshot for international researchers; allowing comparison of manufacturers, types of aids, budgets, and prevalence within a tight geographical domain.

At the time of writing, the Domesday Dataset contains 9,157 purchases from NHS Trusts. Using data from The Office for National Statistics (2011) it can be estimated that the trusts that have responded covers approximately 90% of the UK population. All versions of all datasets are held online for user-access. It is licensed under Open Data Commons Attribution License and meets the requirements for three star linked open data according to Berners-Lee (2010).

The major contribution of this paper is the release of the dataset and the details of the methodology used to construct it. In Section 7, a selection of simple results are presented as demonstrators - concentrating on UK-based issues like the difference in supply based on geographical location (the so-called 'postcode lottery') and the overall spending by the NHS. As a secondary contribution, this work can serve as a blueprint for researchers conducting large-scale studies using FOI legislation, which has already been used for a number of research purposes, most notably in the CHERRI Project by Ellaway et al. (2006), which investigated the process of clinical images for research, but also for environmental research by Jepson et al. (2008).

This paper is organised as follows. Section 2 discusses previous estimates, their scope and any limitations. Sections 3 and 4 detail the data collection methodology and process for synthesising data. Section 5 outlines efforts made in terms of open data and transparency, while Section 6 details the measures taken to preserve the privacy of users. Section 7 gives details of the dataset itself and reports a selection of results that can be obtained from the data and Section 8 gives conclusions and sets out a research agenda including the international expansion of the study.

2. Existing estimates of UK supply

This section examines currently accepted estimates of AAC provision in the UK. The UK is a particularly interesting case study as, due to political activity following the report by Bercow (2008), it had relatively detailed estimation compared to, for example, the United States or Australia.

Although there had been several estimates of the provision of communication devices in the UK, most are based on theoretical estimations of need, or extrapolating from very small populations. This lack of data has been repeatedly stressed as a problem for the area, both academically (see, for example, Baxter et al. (2012a,b)), and politically by Bercow (2008). The most respected estimate is from Gross (2010), which provides two key datapoints:

Year of Purchase	Manufacturer	Model	Num.	Unit Price	Total Price
2006	Liberator	E-Tran Frame	1	£120.00	£120.00
2006	Servox	Digital Electronic Larynx	2	£520.00	£1,040.00
2006	Ablenet	Armstrong Mounting System	1	£190.00	£190.00
2006	Ablenet	Big Mack	6	£84.00	£504.00
2007	Inclusive	Switchit “Bob the Builder”	1	£49.00	£49.00
2007	Cricksoft	Crick USB Switch Box	2	£99.00	£198.00
2007	Sensory Software	Joycable2	1	£49.00	£49.00
2007	Dynavox	Boardmaker	1	£209.00	£209.00
2007	ELO	LCD Touch Monitor	1	£419.00	£419.00
2008	Ablenet	iTalk2 Communication Aid	2	£95.00	£190.00
2008	Attainment Company Inc	Go Talk(unknown type)	4	£130.00	£520.00
2008	Augmentative Communication Inc.	Talking Photo Album	2	£18.91	£37.82

Table 1. Example data from the Domesday Dataset

	2003	2004	2005	2006	2007	2008	2009	2010
Turnover - £	4,040,674	4,377,208	5,109,977	7,337,419	7,561,741	7,805,610	10,463,702	10,199,057
Num. of companies	6	6	7	8	8	7	7	7
Unit sales £1 - £750	1,370	1,518	1,693	1,017	2,777	4,468	4,363	2,546
Num. of companies	5	5	5	5	7	6	6	5
Unit sales £571+	803	832	955	983	1076	1587	2549	1446
Num. of companies	7	7	8	7	7	7	7	6
Software unit sales	n/a	n/a	n/a	n/a	n/a	409	872	740
Num. of companies	n/a	n/a	n/a	n/a	n/a	4	5	4

Table 2. AAC provision information from the eCat Dataset

47. Figures compiled by AAC supplier members of BHTA eCat (British Health Trade Association - electronic Communication and Assistive Technology) show that in 2009 they supplied just over 1,000 hi- tech aids to adults and children in England, with the total spend by schools/colleges, local authorities, the NHS, charities and individuals approximately £3.28m.

48. The NHS accounted for the largest volume of sales and spend (44% of sales and 38% of spend). Local authorities (education and social care) accounted for the next highest volume and spend (18% of sales and 26% of spend), followed by schools and colleges and charities in roughly equal proportions - schools/colleges having higher volume but lower spend.

Although these points have been frequently quoted, it is important to understand the provenance of this information and exactly the definitions used for terms like ‘hi-tech’. Moreover, this data is for England, rather than the whole of the UK.

The information in the communication report was supplied by British Healthcare Trades Association (BHTA) eCat section, which is the loose organisation of companies that supply electronic assistance to the NHS. The membership of BHTA eCat includes many of the major AAC manufacturers, such as Liberator and Toby Churchill, according to British Healthcare Trades Association (2012). The figures were supplied by manufacturers in confidence, to avoid affecting commercial competition (so manufacturers only know the number of devices they claimed, and the total amount claimed by the group). Unfortunately, not only is there no incentive for the manufacturers to be rigorous with these figures, but the position of BHTA eCat is that the subset of BHTA eCat members that supplied the data is confidential. Thus it is not clear what proportion of the industry this represents (for example, Dynavox, one of the biggest suppliers, has been separately confirmed as not included). A final caveat is that the proportional data in point 48 was collected in a different year from the information in point 47 and BHTA eCat have been unable to confirm that the set of companies supplying data is the same in both instances. Clearly, a great deal of care must be taken when extrapolating from this information.

Even with these issues, the eCat Dataset is the highest quality estimate in the field. During the course of this study, BHTA eCat released to the study the more detailed data given in Table 2, allowing analysis at a more granular level.

Given the set of caveats, the eCat Dataset must be interpreted and treated carefully. Although it is extremely valuable as an overview, it is not detailed enough to draw deeper conclusions from, and the methodology (understandably given the

commercial constraints) is not sufficiently public or accountable enough for conclusions to be rigorous. It is recognised that all studies in this area are subject to a certain degree of measurement inaccuracy; later sections discuss approaches to minimise experimental error in the context of the Domesday Dataset. In summary, the eCat Dataset is the current best estimate of AAC use in the UK and both informs the design of the Domesday study, and allows comparison of methodologies.

3. The Domesday Dataset

The central concept of this study is to make a Freedom of Information request to every NHS trust and LEA in the UK and consolidate the data in such a way that not only the data, but also the research methodology, are open to the public and accessible by researchers. Each request required a list of every AAC device the public body had purchased since 2006 including make, model, and year of purchase.

This section gives the methodology used to produce the Domesday Dataset, including the process of designing the study, and collecting and collating the information. In particular it gives details of the process used to normalise the data, including filling in missing information and estimating costs where, for example, companies were no longer trading.

The use of FOI legislation was not the original method of the study; as part of the initial feasibility work for this research, AAC manufacturers were approached with a view to persuading them to release their sales figures. Initial efforts to secure a release of information from manufacturers were, perhaps predictably, unsuccessful in the general case, although some manufacturers were able to assist with validation of the dataset.

One of the most pressing arguments against the use of FOI requests to public bodies, was that the public bodies simply did not have the complete information - as detailed in Section 2 the eCat Dataset showed that only 62% of the sales and 64% of the spend were by public bodies. Even with complete data from the public bodies, researchers would be forced to extrapolate the information, perhaps confirming the trends by means of another research methodology. However, from a political and governance perspective the public body data was the most vital, and it was judged that extrapolating from data as detailed as Table 1 would give extremely accurate results.

There are two assumptions made here - firstly that eCat Dataset's estimation of relative spends (which, should not be skewed by having only a subset of manufacturers submit data) is accurate, and secondly that the figures apply across the UK rather than just England.

3.1. Splitting the data

The Domesday Dataset consists of several individually compiled subsets, divided both by type of public body (NHS for those devices supplied by the healthcare services, and Local Education Authorities (LEAs) for those devices supplied by Social Care or Educational Services), and also geographically (separating out England, Scotland, Wales and Northern Ireland). This split was entirely for the convenience of the data collection process. By separating out the subsets of data it made them easier to compile and validate, and meant that some data could be released while other data was still being requested.

3.2. Design of requests

This section describes the process of designing and 'fine-tuning' the FOI requests made during the study, it goes on to motivate the decision to focus on AAC devices rather than AAC users.

For each of the main types of public body (NHS and LEA), a small number of requests were made with a draft request, as a trial for the main study. Based on the responses from those trusts and feedback from experienced SLT professionals, we modified our request before sending it out to the remaining NHS Trusts.[1]

When designing the requests, significant effort was put into ensuring that the information was not only useful, but also easy for the body to supply, and, importantly, resilient to sources of error. For example, we requested information only from 2006 onwards to reduce the effort required by the public body, and also did not require pricing information, again to reduce the effort required by the public body. We specifically requested only those devices that where purchased by, or on behalf of the trust, to avoid being given a list of those items in the geographic area (and thus having them appear on both NHS and LEA records).

More subtly, trusts would give the number of aids supplied rather than the number of aids purchased, for example Warwickshire PCT[2]. This is of particular interest because, of course, if an aid is supplied to three users in a year - then it is not the case that three users can use it to speak at the same time (recalling the Graeae sisters of Greek mythology). Proponents of using the supply figures may argue that the total number of aids purchased does not account properly for aids used only for assessment, and also that supply figures are more useful in terms of estimating assessment and training rates. The Domesday study takes the view that supply data greatly overstates the number of users at any given time compared to purchase data and is much harder to validate with data from other sources. This issue was consistently raised during the course of the study (a weakness of the study was that during the test phase the language in the Freedom of Information request was ambiguous), and often occurs deep within the data - at least one trust estimated the cost of supplying AAC devices over the year by adding up the value of the equipment within each loan - thus repeatedly counting the value of the same pieces of physical equipment. A strength of the methodology used to construct the Domesday Dataset is that this level of detail can be examined, and validated by any user.

This ‘Graeae issue’ perfectly illustrates the premise that a census of AAC devices will be more accurate than a similarly designed census of AAC users would be. It is much easier for users to be assessed repeatedly, by different bodies, to move houses, and to drop out of the system. Moreover the standards of assessments for devices are not necessarily consistent between, or even within, counties, making it difficult to compare user-focused datasets internationally. We, of course, accept that an equally accurate, equally detailed census of AAC users would be even more rich from a research perspective, particularly to examine things like assessment results for particular disabilities. However, such a dataset would be extremely costly to achieve, and difficult to validate. More fundamentally, it would raise insurmountable privacy concerns, which are discussed in more detail in Section 6. It is the position of this study that extrapolating conclusions about users and AAC provision from the list of devices, is safer, cheaper, and more accurate than extrapolating conclusions from the users.

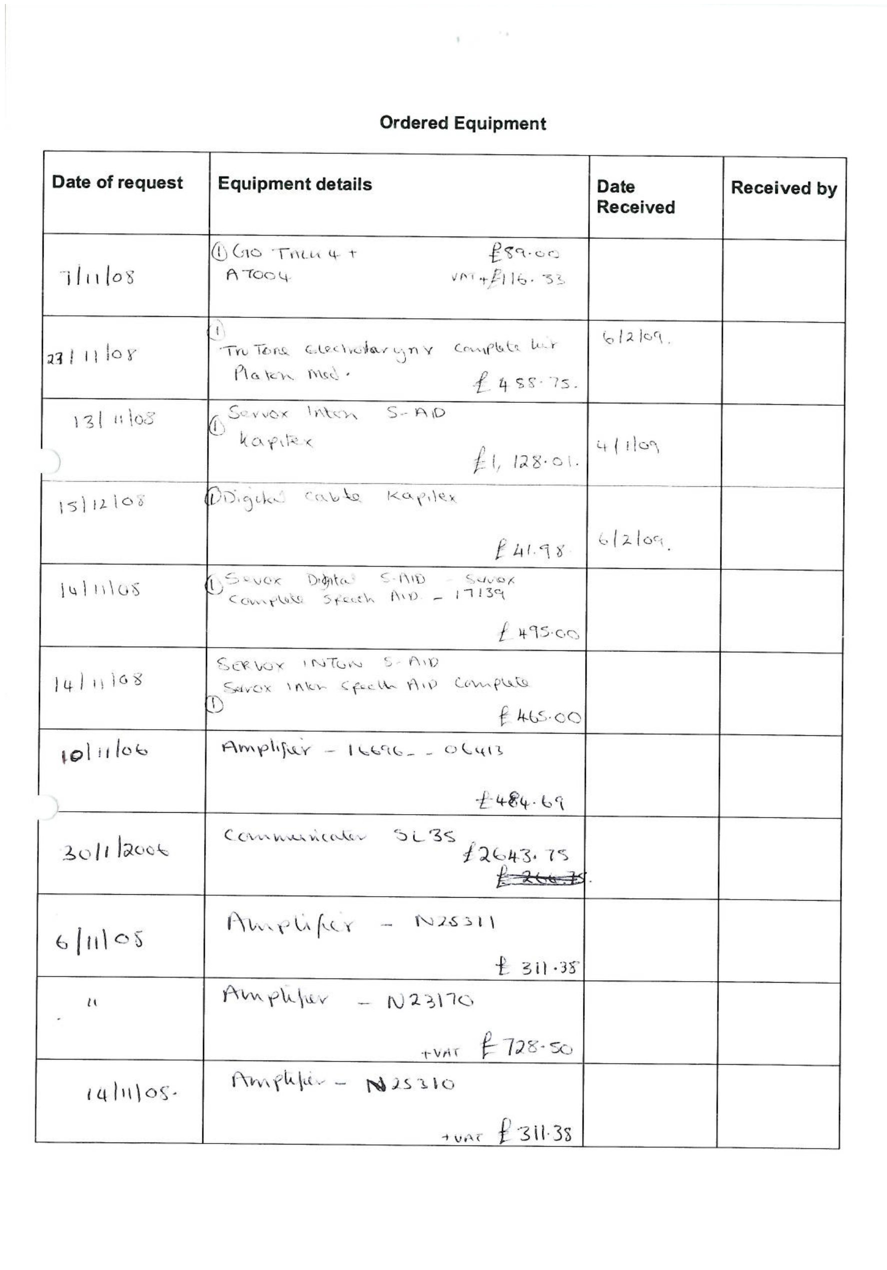
3.3. Following up and collecting the data

This section discusses some aspects of data collection that occurred during the study that may be of particular interest to researchers designing similar projects.

Once each request had been made, a certain amount of follow-up work was often required to recover the data. Public bodies regularly redirected requests to other bodies (such as a Primary Care Trust referring us to a Foundation Trust), in many cases the second trust would refer back to the first trust. In several cases, public bodies cited the Data Protection Act as a reason not to supply the information - when politely challenged on this, they supplied the data. Frequently trusts did not acknowledge the request had arrived at all until contact was made by phone.

Moreover, there are a number of practical or even policy decisions made within Freedom of Information departments that appear to discourage ease of access by members of the public. For example, Leicester City Primary Care Trust replied

[1] An example request can be seen at http://www.cs.rhul.ac.uk/home/joseph/hosted/FOIexample.pdf
[2] http://www.whatdotheyknow.com/request/provision_of_augmentative_and_al_105



Ordered Equipment

Date of request	Equipment details	Date Received	Received by
7/11/08	(1) GO TALK 4 + A T004 £89.00 VAT + £116.33		
23/11/08	(1) Tru Tone Electrolarynx complete kit Platon Med. £458.75.	6/2/09.	
13/11/08	(1) Servox Inton S-AID kapitex £1,128.01.	4/1/09	
15/12/08	(1) Digital cable Kapilex £41.98.	6/2/09.	
14/11/08	(1) Servox Digital S-AID - Servox complete Speech AID - 17139 £495.00		
14/11/08	SERVOX INTON S-AID Servox Inton Speech AID Complete (1) £465.00		
10/11/06	Amplifier - 16696 - - 06413 £484.69		
30/11/2006	Communicator SL35 £2643.75 ~~£264.75~~.		
6/11/05	Amplifier - N25311 £311.38		
"	Amplifier - N23170 +VAT £728.50		
14/11/05.	Amplifier - N25310 +VAT £311.38		

Fig. 1. Example of data provided to the study. This data relates to the Kirklees Primary Care Trust and was sent as a scan.

to an initial request instructing readers to view the pdf file in an attachment. The attachments were a compressed archive and instructions to decompress the archive on a Windows computer. When uncompressed the archive contained a pdf, and that pdf was an acknowledgement that the request had been received and promising a response within the legally mandated 20 working days[3]. These policies appear not to be in the best interests of accessibility. A full evaluation of the Freedom of Information process in these public bodies is outside the scope of this work.

A secondary issue, although a welcome one, was that public bodies would regularly respond with all purchases made by the SLT department - including software, mounting equipment, and non-electronic AAC equipment. Although the study was intended to focus on the set of hi-tech AAC devices rather than the the smaller associated purchases, all the data given by the public bodies is included in the dataset partly to avoid omitting any important data and partly because, having already made the decision that this data would be made entirely open, there was a desire to make sure that the data was useful for all aspects of SLT research. It is important to note that not all public bodies were consistent with the supply of information for these 'low-ticket' items, and researchers are advised to be cautious; perhaps by filtering out certain items by price or item name. A similar unevenness occurred with dates; although only the year of purchase had been requested- it was more effective for many public bodies to give the full date. To include this information the Domesday Dataset includes full-date formats for its information, but those purchases that had only year information are listed as being purchased on the 1st of April of the given year.

Figure 1 shows an example of some of the data we received. Such data was collated into a single dataset and checked over for omissions. At later stages in the process the dataset was checked over for consistency and normalised. This included tasks like using more formal designations for items, such as 'Dynavox MT4', rather than 'MT', separating out items (so 'iPad with proloquo2go' becomes two separate listings) and looking up manufacturers and prices.

In cases where the public body supplied prices the study used that data; however, in the most part prices from manufacturers or resellers were used. In some cases, prices were converted from US dollars to pounds sterling where there was no UK source[4]. In the case of some older items, assistive technology professionals supplied older brochures and price lists. Estimated prices also included VAT at the 2010 rate. Each price estimated includes a reference for the pricing information. Note that this means that there may be some overstating of the amount spent on AAC using this method because many manufacturers reduce their prices for certain clients, and genuinely offer 'multibuy' discounts on this type of medical equipment. Exposing such price negotiation was not a priority for this work.

Moreover there may also be an understating effect on prices because there is no accounting for inflation, nor does the current model account for the fact that, for example, a device might be cheaper in the years following its launch. A more sophisticated model of costings would give a more nuanced view of the data and this is discussed in more detail in Section 7. It is impossible to state strongly enough that the pricing information in this dataset is at a different level of reliability from other data - the Domesday Dataset does not give information about discounted prices paid, warranties, and so on. In many cases 'nearest guesses' were made for model or make, or for products that are no longer sold. For example, when items are listed as "A lightwriter", we took an average price for currently available Lightwriter models as the estimated price.

In some cases public bodies were only able to supply information from a subset of years, largely due to loss of records. Such instances have been recorded in the data.

4. Potential errors in the data

This section gives an overview of possible sources of error in the original data recovered from public bodies. A major priority during the study was to ensure the most accurate data possible - a range of techniques were used to check the information. Firstly, the study's open nature, including all correspondence with the public bodies, allows a deep level of

[3] http://www.whatdotheyknow.com/request/provision_of_augmentative_and_al_42
[4] Prices in original currencies are also included in the dataset

public scrutiny. Potential errors can be introduced into the Domesday Dataset at three main levels: scope; transcription; and interpretation.

At the scope level, the study may have omitted some public bodies or a group of public bodies that supply AAC devices. During this study errors of this type were made and corrected, with the most serious error being that that the NHS service in Northern Ireland was overlooked until a relatively late stage of the process. To allow validation of the scope, a full list of public bodies queried is supplied with the Domesday Dataset, including those bodies that confirmed purchasing no devices in the time period. We discuss those AAC devices purchased outside the scope of public bodies and thus the FOI Act in our future work section.

Transcript level errors are those caused during the transfer from the information provided by a public body to the overall dataset - these include both spelling errors, over-generalisation (listing a 'Lightwriter with mount' as just a 'Lightwriter') and use of the wrong price. Fortunately, this level of error has a relatively low effect and is also easy for casual users to spot because each item in the dataset gives a link to the correspondence that provided it.

At the interpretation level, by which we mean the understanding a given public body took from the request that was made, errors are much harder to detect. Some errors can be simple typographical mistakes, for example Betsi Cadwaladr University Local Health Board included the purchase of 3,267 Techtalks[5].

Interpretation errors that were more difficult to catch included supply/purchase errors as discussed in Section 3.2, which we have been able to deal with to a certain extent with help from manufacturers, and funding issues, where public bodies would give details of all purchases overseen by the SLT team, rather than all purchases made by the trust. This second type of error was often uncovered when both the NHS and LEA bodies reported exactly the same equipment. In general, this type of error poses the most threat to data integrity. Although it is possible for researchers and users to uncover anomalies in the results, only checking by manufacturers or members of the relevant SLT teams can thoroughly validate the results. From spot checks, it is our understanding that overall incidence of these errors is low, and further iterations of the dataset can only improve the integrity of the data.

5. Open data and transparency

A key point about the Domesday Dataset is its open nature. Not only is its goal to create an extremely detailed database of AAC provision, but also to do so in a transparent, reproducible, way - so that the goals, methodology, and results of the study can be suitably critiqued by the community. It is our position that exposing every detail of the methodology is the correct route for the overall benefit to the community. This transparency was greatly aided by www.whatdotheyknow.com (WDTK), which is operated by the charity UK Citizens Online Democracy (2012), and has been previously used in studies like Campbell et al. (2009). WDTK allows members of the public to make Freedom of Information requests in such a way that all correspondence (and data) is available to the public at all times. Each request has a separate URL, which is included as a field in the Domesday Dataset; meaning that any researcher or interested party can examine data and follow the information in the database back to the original publicly made Freedom of Information request. In fact, users are encouraged to do precisely this: a number of manufacturers, although initially unwilling to provide purchase information have been willing to examine the data against their own records and provide corrections. Such manufacturers include Liberator, Toby Churchill, and Claro Software, who between them, account for a large proportion of devices purchased.

This open approach to validation of data naturally lends itself to versioning or forking of the data, which is again actively encouraged. For example, the dataset for Scotland's NHS provision had versions released in May, June, and October of 2012 because not all trusts had replied by the original deadline, more accurate pricing information was available, or because additional information on population size was added. Older versions continue to be hosted so that previous work can be replicated and so that users can follow the evolution of the data.

[5] http://www.whatdotheyknow.com/request/aac_purchases_6#incoming-261833

Purchases	Number of items	Proportion of trusts reporting	Percentage of population covered	Adjusted number of items
England	7205	89%	89%	8,139
Northern Ireland	292	100%	100%	292
Scotland	1146	93%	98%	1,175
Wales	552	100%	100%	513

Table 3. Number of items purchased in dataset, per country, with adjustment for population, 2006-2011

Value	Total	% population covered	%with prices	Estimated NHS Spend	Spend per unit population
England	£4,678,508.34	89%	35.80%	£15,275,601.01	£0.30
Northern Ireland	£249,190.96	100%	56.28%	£442,806.96	£0.25
Scotland	£749,051.67	98%	80.08%	£959,306.44	£0.18
Wales	£458,453.05	100%	78.57%	£583,485.70	£0.19

Table 4. Value of AAC purchases, 2006-2011

Unfortunately, it is not possible to make all correspondence entirely open because regular phone calls were often required to 'chase' trusts for data. In later phases of this study we indicated the use of a phone call with an annotation on the WDTK page for a request; however, this particular data was not necessarily rigorously recorded due to logistical constraints. In general, such phone calls were limited to reminders of the form 'When can I expect a response to my request of the 26th Jan' rather than clarifications of requests or exchange of data.

6. Private data and anonymity

A major concern of the study was to avoid the possibility of anonymity breaches occurring as a result of the release of data.

Privacy issues for AAC users and their families are often undermined by the nature of assistive technology. In particular, the way AAC devices are used can amplify privacy issues; see, for example Reddington and Coles-Kemp (2011); Coles-Kemp et al. (2011). User capabilities result in less empowerment, further complicating the well known and well researched problem of use of anonymised datasets for research, see, for example Hall et al. (2012); O'Hara (2011a,b); O'Keefe (2008). The concern was so key that it informed many of the core decisions made in the study, particularly the approach of focusing on devices rather than users. This design decision was somewhat validated by the respective increase in the integrity of the data as discussed in Section 3.2.

Of course, as part of the Freedom of Information process, all the information collected was first examined by Freedom of Information officers employed by the public bodies to check that no private data had been made public. However, it would be a neglection of duty not to reassess each step of the process. Moreover, even if one were to assume that any one dataset supplied to this study did not compromise anonymity, one must consider the effects of merging them together.

Let us consider the two possible breaches of anonymity that might occur. Firstly, a user could be identifiable as having received their AAC aid from a particular source in a particular year if, for example, they were the only AAC user in a particular community. However, because the Domesday Dataset does not include charity or private purchases (which account for a significant proportion of the overall spend) it is not possible for an individual to be identified as definitely corresponding to a purchase in the Domesday Dataset unless the identifying person also confirms that every other person in the community also does not use such an aid and that no hospitals or other groups have a device of the right type as demonstration models.

The reverse is actually more possible: a user can be shown to have not received their AAC aid from a particular source in a particular year. This is relatively simple, but does not break anonymity unless there is significant other evidence. There is a more insidious threat, which is that if a user claims that they have a device supplied in a particular year by a particular funding authority, and that the Domesday Dataset does not support their claim. In fact - as part of the construction of the study, it is clear that this is not the Domesday Dataset contradicting the user: this is the user contradicting the data. Users who find themselves in such a situation are encouraged to contact the author so that the data can be improved.

7. Results

At the time of writing, the Domesday Dataset contains 9,157 purchases from NHS Trusts. Information from The Office for National Statistics (2011) allows the estimation that the data covers approximately 90% of the UK population. All versions of all datasets are held online for user-access[6].

Although this paper presents a framework for open AAC research rather than major results from the data, it is possible to note some aspects of AAC provision at the high-level.

Firstly, anecdotal evidence of a 'postcode lottery' in AAC provision appears to have some basis in fact. For example, Portsmouth City Teaching PCT, which services a population of two hundred thousand users, supplied details of almost eight hundred purchases. By contrast, the NHS Trusts of County Durham, Hull, Surrey, Trafford, Hertfordshire, Kingston, Mid Essex, Newham, Nottingham City, South Tyneside, Tower Hamlets, Wandsworth, Wirral, City and Hackney, Richmond and Twickenham, purchased no aids at all in the years requested. There are a number of possible explanations for the differences; chiefly that the local educational authorities may be taking up the strain of provision in some areas. The exact circumstances of this postcode lottery are an important target for future research: the trusts reporting no purchases are responsible for a population of over nine million UK residents.

It is also possible to draw conclusions about the effectiveness of industry groups from this data. For example, the current data suggests that BHTA eCat member organisations account for only 33% of items supplied in Wales (this figure is for aids over £750, it drops to 28% for all items in Wales and 25% for all items in Northern Ireland).

Regarding the amount of equipment provided by the NHS, Table 3 shows some details on purchases. Although it has been difficult to extract data from some regions of the UK, principally England, it is possible to use the population figures from Office for National Statistics (2011) to estimate that more ten thousand pieces of AAC equipment were purchased in the UK between 2006 and 2011. Table 4 extends this estimation to overall spend by the NHS. The overall spend figures are somewhat less certain than the number of items purchased, some reasons were given in Section 3, but it is also important to address that the validation of prices in areas like England has been significantly lower than Wales, in which 78% of items have been assigned a price. Table 4 extrapolates from the existing prices to give an estimate but a great deal of weakness is acknowledged. Research to both improve the quality of the pricing and the sophistication of the estimation model would be gratefully received. The estimate by this simple model of NHS spend on AAC equipment from 2006 to 2011 is in the region of seventeen million pounds; however, this is expected to change as more data becomes available. This figure is within the same frame of reference as that of the eCat Dataset, accounting for the proportion of BHTA members within the Domesday Dataset, although we believe a greater proportion of data should be recovered before a detailed comparison is made.

At the qualitative level, one could certainly mine the raw data for interesting circumstances to investigate: elements of the information range from heartwarming (South Tees Hospitals NHS Foundation Trust lists the funders of a laptop used for speech therapy as "Provided by IT Department") to chilling (Norfolk PCT stated "many patients have chosen to be reliant on care staff to interpret their needs via the use of closed questions"[7]).

8. Summary and future work

This work presented the 'Domesday Dataset', which allows researchers to answer fundamental research, policy, and prevalence questions on the landscape of AAC use within the UK. Our work constructed this dataset by means of Freedom of Information (FOI) requests, which asked public bodies (who purchase the majority of AAC devices in the UK) to supply lists of every AAC device purchased since 2006 including make, model, and year of purchase. The resulting repository of information on AAC provision within the UK is more detailed and complete than any seen before. As examples of potential

[6] http://www.cs.rhul.ac.uk/home/joseph/domesday.html
[7] http://www.whatdotheyknow.com/request/92957/response/253036/attach/html/2/Communication%20aids.xls.html

research conclusions that can be drawn from the data, this work presented quantifiable support to the anecdotal evidence of a 'postcode lottery' in AAC provision, some figures on comparative spending within the UK, and an assessment of the coverage of industrial groups for AAC.

The Domesday Dataset is licensed under Open Data Commons Attribution License and meets the requirements for 3 star linked open data according to Berners-Lee (2010). In addition, all elements of the correspondence with public bodies are publicly available online for review allowing validation and checking of the results by interested parties. Furthermore, the depth and openness of the data allow it to be a major resource for AAC researchers in the future. This dataset is not only of use for shaping UK policy and research, but as a snapshot for international researchers - allowing comparison of manufacturers, types of aids, budgets, and prevalence within a tight geographical domain. For example, analysing the proportion of modern AAC devices that have internet connectivity could validate many research goals in next-generation AAC manufacture, see, for example, Patel and Radhakrishnan (2007); Black et al. (2010); Reiter et al. (2009); Reddington and Tintarev (2011).

This work pushes the AAC research agenda in a direction more attractive to larger studies, policy makers, and quantitive research to support the traditionally qualitatively focused field. The range of possibilities for AAC research includes more accurate estimates of populations of AAC users, and levels of AAC use as well as being able to evaluate the success of funding pathways, look into the turnover rate of devices and examine those sectors of the AAC industry that have been most successful at delivering improved functionality to users. At the more fundamental level we hope that this work encourages public debate about where the trade-offs lie in the use of large datasets for research in AAC and the wider intellectual disability field. The oft-cited tension between using information to improve the lives of users and maintaining the privacy and anonymity of users is well studied in medical ethics; but the difficulty of getting such data has previously meant that debate in AAC has been largely abstract. Although the Domesday Dataset largely avoids the issue by focusing on mechanisms rather than users, it demonstrates a workflow for the construction of potentially large, wide-ranging datasets containing a great deal of personal information. It is the author's position that stakeholders at all levels in AAC and intellectual disability should be involved in debate on the levels to which research can, and should, be allowed.

The primary future research goal for the Domesday Dataset is to improve the current dataset by continuing to extract data from the remaining public bodies - a process likely to be greatly aided by widespread use of the existing data by other stakeholders. We also believe that examining the prevalence of AAC internationally would be useful for the AAC community and have a particular interest in repeating the process in the United States, Australia and Sweden. A key part of the study is sustainability, with a view that the process began in 2011 shall be repeated in 2014 to extend the lifetime and use of the resource. Further research goals include examining the explosion of AAC applications for tablet devices to assess their impact on the more traditional AAC manufacturers.

9. Thanks

This research would not have been possible without a large range of contributors, foremost the volunteers and staff at www.whatdotheyknow.com, and the staff at Toby Churchill, Liberator Ltd, and Claro Software. We also thank wholeheartedly David Beukelman for his help assessing the available data on AAC provision internationally. Also, Kate McCallum and Simon Judge for the supply of hard-to-find pricing information and Pam Enderby and the team at Sheffield University, along with Tom Bamber at Loughborough University for many useful conversations. We thank David Morgan, Anna Reeves, and Sandra Hartley for a great deal of information and encouragement and Liz Panton for the expert reviewing of the requests from an SLT perspective. The advice of Michael Hausenblas with aspects of the open data and transparency were invaluable and the contributions of Lizzie Coles-Kemp, Nava Tintarev, Amy Brown, Sarah Moffat, and John Carter in shaping the research were considerable. Most importantly we thank those FOI officers and SLT leads that responded to the study efficiently and effectively - this work absolutely wouldn't be possible without you.

References

S. Baxter, et al. (2012a). ‘Barriers and facilitators to the use of high-technology augmentative and alternative communication devices: a systematic review and qualitative synthesis’. International Journal of Language & Communication Disorders .

S. Baxter, et al. (2012b). ‘Interventions Using High-Technology Communication Devices: A State of the Art Review’. Folia Phoniatrica et Logopaedica 64(3):137–144.

J. Bercow (2008). The Bercow Report: A review of services for children and young people (0-19) with speech, language and communication needs. Department for Children, Schools and Families.

T. Berners-Lee (2010). ‘Linked Data’. Personal website (‘http://www.w3.org/DesignIssues/LinkedData.html’).

P. Birkinshaw (2010). Freedom of Information: the Law, the Practice and the Ideal. Cambridge University Press.

R. Black, et al. (2010). ‘Using NLG and sensors to support personal narrative for children with complex communication needs’. In Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies, pp. 1–9. Association for Computational Linguistics.

British Healthcare Trades Association (2012). ‘BHTA, http://www.bhta.net/’.

D. Budgen, et al. (2005). ‘Managing Healthcare Information: The Role of the Broker’. Studies in health technology and informatics 112:316.

M. Campbell, et al. (2009). ‘A survey of allotment waiting lists in England’. Transition Town West Kirby with National Society of Allotment and Leisure Gardeners. UK .

L. Coles-Kemp, et al. (2011). ‘Looking at clouds from both sides: The advantages and disadvantages of placing personal narratives in the cloud’. Information Security Technical Report .

R. Ellaway, et al. (2006). ‘Clinical recordings for academic non-clinical settings’. Joint Information Systems Committee. Retrieved September 17:2010.

J. Gross (2010). ‘Augmentative and alternative communication: a report on provision for children and young people in England’.

W. Hall, et al. (2012). Open data and charities. Nominet Trust.

P. Jepson, et al. (2008). ‘Investigation of the common dolphin mass stranding event in Cornwall, 9th June 2008’. UK Cetacean Strandings Investigation Programme, London .

Office for National Statistics (2011). ‘Primary Care Organisation Population Estimates’. Office for National Statistics .

K. O’Hara (2011a). ‘Interim report on privacy and transparency with respect to the release of crime data: open letter to Francis Maude, Minister for the Cabinet Office’ .

K. O’Hara (2011b). ‘Transparent Government, Not Transparent Citizens’. Report on Privacy and Transparency for the Cabinet Office .

C. O’Keefe (2008). ‘Privacy and the use of health data-reducing disclosure risk’. electronic Journal of Health Informatics 3(1):e5.

R. Patel & R. Radhakrishnan (2007). ‘Enhancing Access to Situational Vocabulary by Leveraging Geographic Context’. Assistive Technology Outcomes and Benefits p. 99.

J. Reddington & L. Coles-Kemp (2011). ‘Trap Hunting: Finding Personal Data Management Issues in Next Generation AAC Devices’. Proceedings of the second workshop on speech and language processing for assistive technologies pp. 32–42.

J. Reddington & N. Tintarev (2011). ‘Automatically generating stories from sensor data’. In Proceedings of the 15th international conference on Intelligent user interfaces, pp. 407–410. ACM.

E. Reiter, et al. (2009). ‘Using NLG to help language-impaired users tell stories and participate in social dialogues’. In European Workshop on Natural Language Generation (ENLG-09).

J. Strobl, et al. (2000). ‘Data protection legislation: interpretation and barriers to research’. BMJ: British Medical Journal 321(7265):890.

UK Citizens Online Democracy (2012). ‘WhatDoTheyKnow.com, http://www.whatdotheyknow.com/’.

B. Worthy & G. Bourke (2011). ‘The use of FOI by Parliamentarians and the Impact of FOI on Parliament’ .

C. Zangari, et al. (1994). ‘Augmentative and alternative communication: An historic perspective’. Augmentative and Alternative Communication 10(1):27–59.