

The Internet Directory is NOT Primarily a Technology Issue

Position Paper by D W Chadwick

Position

The main reasons why we do not yet have a fully operational Internet Directory service, whether it be based on X.500 or on Whois++ or on 'this-great-new-protocol-that-I-have-just-invented' :-), are primarily managerial, administrative and legal, and are NOT technical.

Rationale

This position is drawn from personal experience of working with directories over the last decade, as well as from recently published research. The latter comprises a study conducted in the UK by Paul Barker [1], and a series of X.500 case studies [2] conducted by myself over the last couple of years.

Building an integrated electronic directory service within an organisation requires a lot of senior management commitment - commitment to provide resources in terms of staff time, computer hardware and software, and commitment to coerce, cajole or force other departments in an organisation to release 'their' data for publication in someone else's electronic directory service. Some managers get very possessive over 'their' information. So top management support is often needed in order to get access to the data. Coupled to this is the spectre of information confidentiality. It is sometimes argued that publication of organisational telephone numbers and Email addresses in an electronic directory is breaching the personal privacy of the employees. So this is another reason for not being able to gain access to the data for the directory. To quote one example, a research organisation in Germany did at one time have all its staff in an X.500 electronic directory connected to the Internet, but this had to be removed following a workers council decision that required the organisation to obtain written permission from every employee before their details could be entered into the directory. The administrative hassle of this was seen to outweigh the benefits, and so the data was removed. Another example, quoted in [2], concerns a Scottish University, where senior managers would not release data from the personnel department for publication in an X.500 directory, as it was seen to be breaching confidentiality. The decision was eventually reversed after several years of negotiation, but by this time the impetus to install the directory was lost.

Once the necessary permissions are received to load the data into the directory, your data management troubles are not over. On the contrary they could be just starting. Reference [2] found that many sites had data consistency problems. The master data sources such as the HR database, the PABX and the Email address file were often found to be inconsistent, with names spelt differently, initials and titles either missing or in reversed order, and some staff completely absent from one or other data source. It can literally take several man weeks of effort just to sort out the data inconsistencies, with manual intervention usually being the only effective way. Most of the sites studied significantly underestimated this problem. Reference [1] quotes an average time of 5-6 man weeks of effort, with the worst case taking 4 months. Finally software tools will have to be purchased or built, and procedures written, so that the

data can be frequently and automatically updated. For these reasons some sites studied in [1] chose to opt out from having an integrated directory service, and instead chose to have separate telephone and Email sequential files, containing different sets of names.

To summarise, it takes a lot of time, patience and resources to collect all of the necessary information, load it into a directory, and keep it up to date, and so senior management support is essential.

Coupled with this we must realise that users have very stringent demands on the quality of data that is kept in an electronic directory. It must be 100% accurate, it must be current, and it must be complete. Otherwise the directory will soon fall into disrepute. *Accuracy* refers to information such as names and addresses being correct, *currency* refers to data being continually updated (at least daily for human consumption, much more frequently for computer based applications) and *complete* refers to the breadth of data that is stored. Contrast this with the Web which is enjoying phenomenal success at the current time, but which has few (if any) of these data quality requirements placed on it. When I query the Web, I might be very happy to find any information about topic z. It does not matter too much if a particular document is 3 years old, I can still read it and gain valuable information from it. Nor does it matter if I am unable to retrieve all the information there is about topic z, I may be perfectly happy to pick up less than 10% of the global published works. But consider approaching the Web with a very specific query, for which there is only one correct answer e.g. I want to find the population of Santa Monica, or I want to access an online copy of the Merchant of Venice, or what is the current price of product x, or what is the ISBN of 'Understanding X.500'[3] etc. A Web user might find it very frustrating trying to track down this sort of specific information in the Web. But these are precisely the sorts of queries that are levelled at a directory service. I need to know the telephone number of Jo Davis in Digital, who works somewhere in the US, or what is the Email address of Thomas Wolchenizia at the University of Prague. Surprise, surprise, most queries of this type when levelled against the current Internet X.500 directory service fail, so users tend to discount the technology as not being up to job, when in fact the failure has nothing to do with the technology, and everything to do with the management of the data within the directory. Management can choose precisely what information they publish in the Web, but with the directory they have much less freedom of choice. An Internet directory which does not hold communications related data is no directory at all.

By contrast, there are a number of excellent well managed X.500 directory sites around the globe that will always give you the data you require. For example, Salford University holds the names of every undergraduate and postgraduate student and every member of staff in its X.500 directory, and it is updated nightly. It holds telephone numbers, fax numbers and Email addresses. Other good quality data sites that I have queried in the past include NASA and the University of Michigan. But these are few in number compared to those sites that do not have a good quality globally connected directory service. Why is this?

Part of the reason is that being part of a global directory service is very difficult for management to justify in terms of a simple cost benefit analysis. The immediate cost of implementation is high, but the immediate benefit is seen to be comparatively low. Reference [1] found that approximately half the universities in the UK were not prepared to countenance this cost, and instead opted for a simple local directory service that fulfilled the most pressing requirements of their local users, particularly local Email address lookup. One of its major

findings was that a local directory service was seen to solve most of the directory service problems of a site, so there was no desire to foot the extra cost of buying into a global directory service. A few sites were not even prepared to fund the implementation of a local directory service. Even when a directory service is fully functioning and well utilised by staff, the number of daily directory queries can be orders of magnitude less than the number of phone calls made, or the number of Email messages sent. Consequently an organisation can operate without an electronic directory service, but it cannot operate without access to the telephone, and increasingly, Email. Hence the reluctance, by some universities at least, to invest their scarce resources in a 'nice to have' part of the infrastructure. If the organisation has lived without an electronic directory for the last 20 years, surely it can live without it for the next year or two until the cost of implementation is reduced. The hidden costs of staff time wasted trying to find addressing information, are rarely taken into the cost benefit equation.

Another reason concerns commercial confidentiality. Commercial organisations may successfully implement an internal directory service, and they may even connect it to the Internet directory service, but they certainly will not make *their* information accessible to the global community. They are only interested in retrieving information from the directory, not contributing to it. The spectre of confidentiality rears its frightening head once again. (It is not the purpose of this paper to embarrass those companies who have already done this using X.500 technology, by listing their names in public, but if you want to know who they are, look them up in NameFLOW-Paradise. They are immediately identifiable as you are only be able to retrieve a stub entry giving the name, address and phone number of the corporate headquarters. You wont find any staff details.) Again, this often leads to criticism of the underlying technology, when in fact it is purely a result of a management decision.

For an Internet directory service to be successful, a huge critical mass of data is required. In my experience users can tolerate poor network performance, if the chances of success are high (>90%), i.e. a user has a greater than 90% chance of finding person X in the Internet directory. The NameFLOW-Paradise service currently has approaching two million entries in X.500, but this is not seen to be sufficient to guarantee a reasonable chance of success. The chances of finding the details of more than say 10% of the people at this conference in the NameFLOW-Paradise service are probably less than 50/50. So if I need to find YOUR details, where do I look? This situation is made more frustrating when one considers that we now have competing and incompatible directory technologies. This will inhibit the creation of a critical mass of users in any single directory. If we are to accept competing directory technologies on the Internet, then some form of gatewaying between them is absolutely essential.

In this position paper I hope to have persuaded you that many of the inhibitors to a global Internet directory service are primarily managerial, administrative, and to a lesser extent legal, and are not as a result of poor or inappropriate technology. For these reasons, I do not think that new protocol developments such as Whois++ will significantly affect the overall growth of an Internet directory service, unless it can significantly reduce the cost of entry (and given that Quipu was originally supplied free of charge, this does not seem probable). However, the Nomenclator pilot [4] might prove to be more successful, because this is attempting to interconnect already existing CCSO databases containing directory information. Hence many of the management, legal and administrative problems outlined in this position paper have already been overcome.

We all look forward to the day when we will be able to find each others' details in the Internet directory service, whatever its underlying technology. Let's hope that we will be able to do that before we retire.

References

[1] Barker, P. "White Pages Directory Services in the UK Academic Community", December 1995, available by anonymous FTP from [cs.ucl.ac.uk/dirpilot/wpds.ps](ftp://cs.ucl.ac.uk/dirpilot/wpds.ps) or <http://www.bath.ac.uk/~ccsap/Directory/Docs/x5usage/x5usage.html>

[2] Chadwick, D.W. "Important lessons derived from X.500 Case Studies", IEEE Networks, March/April 1996

[2] Chadwick, D.W. "Understanding X.500", Chapman and Hall, July 1994, ISBN 0-412-43020-7

[4] Nomenclator details available from <http://www.cs.att.com/csrc/nomen/nomenclator.html>