GRESHAM COLLEGE

15 JANUARY 2019

THE SEARCH FOR MEANING ON THE WEB:
THE SEMANTIC WEB AND MANAGING A LIFETIME'S INFORMATION

PROFESSOR YORICK WILKS

We cannot separate the modern history of AI from that of the WWW which is its main interface to the public. The WWW is perhaps a good example of an unintended consequence, something more common in science and technology than you might guess. Saying that assumes the notions of surprise and serendipity: the latter is very close to that of things being *designed for one purpose but actually found to work for another*, as shown by the classic case of Post-it notes: made with a weak glue perfect for the notes but little good as the real glue it was meant to be. Richard Gregory claimed in the 1960s that the brain was an engine designed largely for seeing but which has had its "seeing mechanisms" adapted later to yield our language faculty.

Some of the most striking developments on the Internet in recent years are simple surprises in that no one at the beginning of the 1990s could possibly have predicted:

* The sheer scale of media sharing such as *YouTube*
* Autobiographical social networks of personal data: *Facebook*
* The reduction ad absurdum of instant communication: *Twitter*
* Mass texting for communication and the *telegraphese* of abbreviations it created.

The last is extraordinary given that text was originally a maintenance facility on phones for engineers. In the early phases of successful automatic speech research, it was taken for granted that speech recognition would oust all written forms. As US military funders used to put it "Generals don't type".

Unforeseen novelties created on top of the WWW include;

- The return to the timeshared mainframe model (now called the Cloud) from storing data on a personal machine
- The Web takeover of the phone system (initially by *Skype*)
- Peer-to-peer sharing of data was totally unforeseen, and occupies the WWW all night, like dreaming in humans!
- The new paradigm of "crowd science", such as *Wikiproteins* and also crowd knowledge creations such as *Wikipedia*

It is not always understood that many of the features we associate with the WWW were available on the earlier forms of the Internet when it was still dominated by its origins as a military-academic communication system set up by DARPA in the US. There was already the ability to shift large files round the Internet from site to site: the protocol was not the current web protocol (called http, which you see in all WWW addresses) but ftp, the "file transfer protocol". I can remember Marvin Minsky at MIT sending out the whole of a draft book to everyone on the Internet in 1972, the first time such a thing had been done. The Internet itself—or the ARPAnet as it was originally called--- was the brainchild of J Robert Licklider and designed for US military needs: how to get data and information seamlessly across a network of US government computers. It was, in computer science terms, a "packet switching network": the breaking up of information into tiny packets of bits and sending them off across

a network of cables or microwaves and hoping they would reassemble properly into documents at the destination. That was the ARPAnet, and it was decided early on to include the great AI research laboratories such as Stanford and MIT in its network. The ARPAnet thus hugely accelerated the progress of AI research: soon after email was invented in the late 1960s it drew the US AI research community into a single whole, though even then much of the email traffic was already about personal needs and assignations, just as now.

The ARPAnet later became the Internet and was made world-wide and "civilianized" but before that, and long before the arrival of Berners-Lee's WWW which was carried by the Internet, there were facilities such as *Usenet*, very like the later WWW. Usenet was a system of peer-to-peer file transfer with no central organization which supported a huge range of user groups that were open to anyone, and which sent images as well as text: all interests were catered for right down to alt.sex.com. Blogging certainly started there, rather than on the later WWW, and many people set up file sites designed to lure users into reading personal diaries.

The World Wide Web

The arrival of the WWW in 1992 changed everything: Berners-Lee's original vision, conceived at an international physics laboratory in Geneva, was essentially to give every document a physicist might want a name (later called its URL—its Universal Resource Locator) so it could be seen from anywhere in the world, regardless of where it was stored. It stimulated the creation of new ways to search the names of web-servers and their files, none of which were part of the original WWW design itself. Berners-Lee, when he started the WWW, just had a list of servers holding documents until that number became too large to be manageable.

The Semantic Web

I started this lecture with the idea of serendipity because the WWW that Berners-Lee gave the world in 1992 was not what he had originally intended: what one might call Berners-Lee's first great idea was that of a *Semantic Web* (SW), which I will now describe briefly. The WWW we actually got was serendipitous, like post-it-notes, because it did not have the key property that he was looking for in the SW, which was to express the *meaning* of what it contained. The SW is no longer simply an aspiration but is now a serious research subject on both sides of the Atlantic, with its own conferences and journal. Even though it may not yet exist as a single demonstrable entity, in the way the WWW itself plainly does, it exists in part and is a topic for research and about which fundamental questions can be asked as to its representations and their meanings.

The WWW is like TV: it has millions of documents and images on it but it doesn't know what is in them, just as a TV set does not know what it is showing—it just puts it on the screen. The SW, that Berners-Lee originally wanted, was system of documents and images where the system itself knew what they contained and *could reason about their content*. The SW therefore is much more like AI than the WWW we actually have, because it could show intelligence.

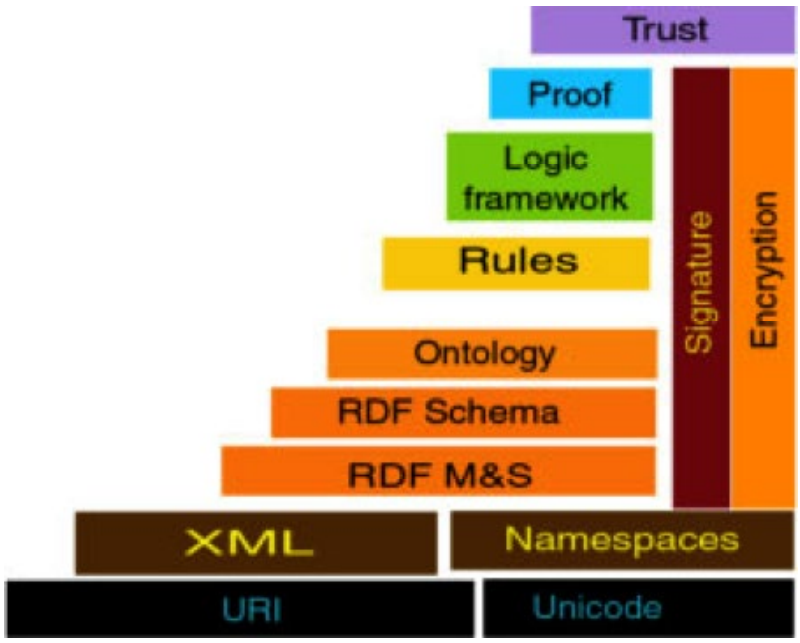
In his paper in the *Scientific American* in 2001 that set out his design for the SW, Berners-Lee described making an appointment for an elderly relative, a task that would require planning and access to both the doctor's and relative's diaries and synchronizing them. This kind of planning behaviour was at the heart of what I have been calling classical AI. If one looks at the SW diagram from the original *Scientific American* paper (see below) it shows the SW as a stack of "levels of representation", which just means symbols that express the meaning of other symbols. That is very like what dictionaries do for us: we look up a word and get other words telling us what the word means. In the SW stack the "lowest" level is the base in the Internet that carries the WWW, and would carry the SW. The tendency has always been to look at the upper levels of this figure: rules, logic framework and proof, and it is these that have caused both critics and admirers of the SW to say that it is the classic AI project by another name. But if one looks at the lower levels one finds *Namespaces* and *XML*, which are all the products of what I have called NLP (Natural Language Processing) obtained by annotations of texts by a range of NLP technologies. But for now I will take *annotation* to be what scholars in the humanities have always done with texts, adding *marks* to them to show their content.

So, "markup languages" (and XML is one such language) grew up in humanities departments as ways of marking,

say, that a name in a text was a proper name (perhaps of an Egyptian Pharaoh). This technique was taken over by NLP researchers to mark texts with their grammar, such as marking a particular noun as a noun. That sounds obvious, but later we shall see that this was done so that a program could *learn* a grammar if it was shown enough examples of what a noun was, and that was done by putting marks into a text after a noun like "Ptolemy", such as <proper noun>.

It is important to remember that available information for science, business and everyday life still exists overwhelmingly as text: 85% of business data still exists as unstructured data (what text is now technically called). So, too, of course does the WWW, though the proportion of it that is text (as opposed to diagrams, photos, videos and tables) is falling. And how could the WWW become the SW, one might ask, except by information being extracted from natural text and stored in some other form along with it. NLP in the 1990s created technologies that could do that to text: to store *within the text* indications of what its parts meant and what their structure was. The inspiration for this didn't come only from the humanities but also from printing technology: an early side-interest in AI laboratories , and computer science more generally, in the 1970s was printing documents on the computers then available. This was done by systems for inserting marks into a document's text, indicating how what followed the mark should be printed, such as putting it in italics or in a larger font.

The most sophisticated of these was a printing language called *Tex*, (later, *Latex*), invented by Donald Knuth at Stanford, originally to manage the look of his own work on the page, but which became the main tool mathematicians used to layout and print their formulas in their papers. At about the same time humanities scholars developed *TEI*, the Text Encoding Initiative, for the scholarly mark up of text which then developed into *SGML*, a General Markup Language, and then into *HTML*, the mark up that has become, with its successor *XML*, the Extensible Markup Language, the standard for encoding all webpages on the WWW. Later on, special markup languages were developed for Time (*TimeML*), Speech (*VoiceML*) and a plethora of other special areas. They all share the idea captured by those in-text printing commands: that you insert special marks into text that are not part of the text itself, but which tell you what to do with the text, and how to interpret it, or what it means. These "marked-up" pages, created by the page designer, underlie every webpage you see: they are the bones beneath its easy on the eye flesh.



The structure of the Semantic Web

On this view, the SW can be seen as a conversion from the WWW of texts by means of an annotation process of increasing depth, one that projects notions of meaning up the classic SW diagram from the bottom. The philosopher of science Richard Braithwaite once wrote an influential book on how scientific theories obtain the semantic interpretation of "high level" abstract entities (such as neutrinos or bosons in physics) from low level

data in observations. He named the process one of *semantic ascent* up a hierarchically-ordered scientific theory. The view of the SW under discussion here, which sees NLP as providing its foundational processes, bears a striking resemblance to that view of scientific theories in general.

The SW and Data Bases

Berners-Lee's views of what the real SW web was to be ----- the one he really wanted and not just the WWW-----developed with the availability of very large data bases, into what is often called the era of "big data". Some now see databases as the core of the SW: databases, the meanings of whose features are kept constant and trustworthy by a cadre of guardians of their integrity, where "features' just means the indicators, always in English words, of what the database stores, such as BIRTHDATE. This integrity is important because words change their meanings with time: we still "dial numbers" when we phone, even though that no longer means the finger actions it did a few decades ago. So not even number-associated concepts are safe from time. The word "Guardian" here expresses a view that (computer) scientists can preserve us from changes of meaning, so that the content or interpretation of data cannot change.

Spärck Jones' view of AI owes something to a theory of the American philosopher Hilary Putnam in the 1970s: he believed scientists were *guardians of meaning*, since only they know the true chemical nature of, say, molybdenum and how it differs from the superficially similar metal aluminium. So, said Putnam, only scientists *really knew* what those words mean. For him it was also essential that the scientists did not allow the criteria of meaning (for being, say, aluminium as opposed to being molybdenum) to leak out to the general public, lest they then became subject to change. Thus, for Putnam, only scientists knew the distinguishing criteria for separating water and deuterium dioxide (heavy water) which seem the same to most of the population but are not. I personally do not think such a separation (of popular meaning from that known to expert guardians) makes sense, quite aside from how profoundly undemocratic it is.

The heart of the issue about content can also be put as follows: we have the feeling that we locate things in our minds by *content,* that we search by what something means. The SW represents an attempt to realize that feeling in the world of information search. Sparck Jones argued that whenever that desire is realized, we later find that some statistical method, that makes no reference to meaning, can do the job better. That is the "Guardian of content" issue at the SW's heart and it remains unresolved: with head and hearts at odds with each other.

What did Berners-Lee intend when he began all this?

The conventional view is that Berners--Lee first set up a web of documents (the WWW) and then shifted his interests towards what we now call the SW. I believe this is not correct and that there is an element of later serendipity about the WWW as we have it, and that he always intended something like the SW. Here is his own account in 1998 of what he initially did in Geneva:

> *"…in 1989, while working at the European Particle Physics Laboratory, I proposed that a global hypertext space be created in which any network-accessible information could be referred to by a single "Universal Document Identifier"……I wrote in 1990 a program called "Worldwide Web", a point and click hypertext editor which ran on the "NeXT" machine. This, together with the first Web server, I released to the High Energy Physics community at first, and to the hypertext and NeXT communities in the summer of 1991 ……The great need for information about information, to help us categorize, sort, pay for, own information is driving the design of languages for the web designed for processing by machines, rather than people. The web of human-readable documents is being merged with a web of machine-understandable data. The potential of the mixture of humans and machines working together and communicating through the web could be immense."*

Berners-Lee's historic innovation in the WWW---and it is important to remember also the key roles of others such as Andreessen's MOSAIC browser and hypertext itself, an idea of Ted Nelson's----was that of a document's *unique identifier*. The key issue is :identifier of what? In his recollections, quoted above, he uses the phrase "Unique Document Identifier" (UDI) as a simple name—one which has disappeared from the lexicon and does not appear in Wikipedia, for example. It was soon replaced with the, still familiar, Universal Resource Locator (URL) which

is normally taken as the unique name of a document. All this leads naturally to the WWW as we know it, as a web of documents.

Why is this of importance? Well, the reminiscence above can also be read another way. Berners-Lee refers explicitly to: "The great need for information about information…. is driving the design of languages for the web designed for processing by machines, rather than people. The web of *human-readable documents is being merged with a web of machine-understandable data*. The potential of the mixture of humans and machines working together and communicating through the web could be immense".

This is precisely the spirit of the original SW document he wrote later with others in the *Scientific American,* but even in 1989 he added a diagram in which you can see the lines from documents pointing to objects in the real world. That is not just a web of documents, such as the WWW, but also of *things*; which is to say data, or data standing for things, a matter much closer to his heart, as a database specialist, than language in documents. That, I believe, is the evidence for his real view of the SW, even early on: the web of data, data for machines rather than human readers.

So, Unique Resource Locators (URLs) became the piece of web technology that got into everyday speech, being the name for a document that is, say, a book and which will pick out or point to that book's content, stored on the WWW. That name does not pick out in the world a unique real book, an object weighing 300grams---in the way "Tim Berners-Lee" *does* pick out a unique object, namely him----because there may be thousands of copies of any book, and only the RFID tag (derived from *Radio Frequency Identifier*) that the bookseller puts in it to stop shoplifting identifies a particular hard copy of it. These ambiguities over the names of general classes of things, versus particular individuals, are familiar in the history of philosophy and meaning, and it is illusory to think that the SW has actually solved two millennia-old problems about names and how they connect to the world of things. But it has given us a vast working example of how a system of names and things can function practically, which is another example of refuting the "you can't do that" arguments we encountered earlier from logicians.

The origins of the WWW are in Berners-Lee's databases for physicists and unconnected to AI. But its sophisticated search capacities developed from within AI and it is, I believe, moving inevitably if slowly towards some form of SW, one where representations of knowledge and meaning will be central, and those, as we have seen have always been at the heart of AI. The gigantic computer structures of all medical and scientific knowledge on the web are now indispensable for researchers, and their structure and interpretation has come from decades of work in the mainline AI tradition. Google's own *Knowledge Graph* is essentially an application of the SW idea.

Another major point about the SW project and its relevance to AI is the philosophical issue of *grounding*: how do our symbols in language or in AI programs link to real objects in the world that *ground their meaning*. Wittgenstein's treatment of the problem assumed that it was wrong just to think of this as being about how the word "apple" links to a real apple. He suggested it is rather a question of how the vast web of language with all its links from word to word connects to the whole world. The WWW is just a network of links from words to documents and pictures and we can ask now if anything in it has any real meaning. What we call URLs end up pointing to yet more symbols, to more documents; so how can we ever escape from that world of symbols, to some firm *ground*? Perhaps the success of the web, and even more of the nascent SW, is that we do not have to have that escape for the whole to be useful and to give us the answers we need! In some sense, its very functionality shows that the ancient philosophical problem---of attaching words to things---- does not stop our understanding of the world of symbols. It has always been posed in terms of words such as "apple" but our real lives are full of quite different words such as "Germany", "courage", "love", "lies", "procrastination" and "Hamlet" which point at nothing at all that is concrete in the world, and yet we cope.

We shall now take a step back and say something about the notion of a computer program, on which all AI ultimately rests, after which we shall move on to particular areas of AI.

The Companion as A Way into The Web

As late as 1996 the Investors' Business Daily could write in an editorial: *A senior VP at International Data Corp. predicts widespread consumer disillusion with the Internet: "I also believe the Internet and World Wide Web phenomenon will go from an intoxication stage to a hangover stage during the next two years. About 20% of the large firms will jump off the Web because the consumer demographics are not developing quickly enough*". A few years later the Guardian reported a similarly misguided complainant who said that '… the internet is killing their trade because customers…seem to prefer an electronic serf with limitless memory and no conversation.' It is above all this lack of conversation with the Internet that the Companion might remedy. I assume in what follows that any such Companion, such as the Senior Companion of the last lecture, can be seen as a prime exemplar and incarnation of all major AI technologies brought together in one place, rather as Wagner called opera the *Gesamtkunstwerk*, the total art work.

I introduced Companions in the last lecture as specialized computer agents that could converse and assist needy groups of people such as the old. That help will almost certainly go way beyond helping organize their lives and memories, but will extend to interacting with the electronic world outside for them as web agents. Web agents already have to make deals and transactions of all sorts with each other, and learn to trust each other in, for example, the world of banking where agents in their thousands clinch money transactions between financial institutions.

These activities may come to require exactly the kind of computer agency a Companion will be able to offer to a person in dealing with the web as it becomes more complex. To put this very simply: the web may become unusable for non-experts unless we have Companion-like agents to manage its complexity for us. One reason for this is that there is now just too much data in our lives to cope with, both external data we need to access as well as all the data we generate ourselves, such as thousands of photographs of ourselves and our meals. The other reason is that people are not very good at web search, but do not know this: they can input names and sex and holidays into Google and get something useful back, but hardly anyone ever looks beyond the first page of what a search engine returns, and they rarely filter their search results --- keeping in certain terms and dropping others—in ways that web search allows but non-specialists rarely know about.

So, the web may have to develop more human-like characteristics at its human interface to survive as a usable resource and technology: just locating a particular individual on the web, when a majority of the world's population have a web presence, will become far more difficult and time consuming than it is now. Companions will then be needed by everyone, not simply the old, the young and the disabled. It may become impossible to conceive of the web without some kind of a human face that renders it personal, a term which has sometimes become distorted to mean "getting only the news you agree with", which has in turn become part of the problem. One could put this by saying that AI and web technologies have created these problems, and they are going to have to solve them.

Discretion in a Companion

I discussed earlier the 19C idea of a Victorian Lady's Companion and noted that one of her key features was *discretion*, and it may be worth thinking a little more about that in today's terms. How much might one want to keep secret from one's Companion if you did not trust their discretion, their commitment to just you as the owner? The household presence of ALEXA is already causing such concerns, with users becoming aware that ALEXA is listening to domestic conversations and almost certainly reporting them centrally to its maker, ostensibly for "learning and improvement" purposes. A US hotel chain has just announced that ALEXA will be installed in all its rooms and replace concierge services such as luggage and room service. But will future guests perhaps be uneasy at the "discretion" of ALEXA concerning conversations in the room it will almost certainly hear, perhaps without any indication it is on?

More generally, will future Companions share their knowledge behind the backs of their owners, in the way that Victorian domestic servants were generally believed to do, but Lady's Companions were supposed not to? How confident will one feel, in the midst of a family feud, that the instruction "Don't tell Billy this" to an artificial Companion will always be obeyed? Much of this will be moot because Companions will normally deal only with

one person—which is what makes their speech recognition problem so much easier, as I noted—they will be trained for a single speaker---except when, say, making phone calls to an official, a friend or a restaurant. Google recently demonstrated an automated call to reserve a restaurant table, but this has been demonstrated for some twenty years now. Like much AI, the gestation period to market is much slower than many believe.

The notion of a stored fact that must *not* be disclosed is simple to code, but the problem is wider in that the same fact must, to preserve the secret, not take part in inference processes either. If it is a secret that Tom is really a Russian, then the Companion should not do inferences such as [IF X is of nationality Y THEN X will normally speak Y] and suddenly come out with "I assumed Tom could speak Russian", which would rather give the game away.

On the other hand, a Companion gossiping behind an owner's back could be a positive development that we might encourage: imagine the old person in a care home, too shy to approach another for a lunch together. This would be something best settled between their Companions, each knowing the tastes and habits of their owner, to whom the "date" could be presented as a *fait accompli*. Again, many Companion-to-Companion interactions will be between an individual's Companion and some form of "public Companion", such as one that takes restaurant bookings based on a user's tastes; or at a hospital where a hospital-Companion would triage incoming patients, who may not be articulate about their condition, on the basis of detailed knowledge of the user's medical records revealed by their Companion. When traveling, this Companion-to-Companion interaction in, say, a hospital could also combine with translation where the respective Companions worked out how to communicate across a language barrier.

In all these cases, Companion-to-Companion communication could be of obvious benefit to a user even if confidential information was at risk of disclosure: the user might have said "Never tell anyone I'm HIV positive" but in the hospital environment that constraint should obviously be overridden and the user's condition revealed. One could say at this point that secrets may be relative to a situation and that there may be nothing more complex in a Companion's guardianship of secrets than there is in explicit restrictions one could give to human hearers. In some situations, a Companion might owe a public loyalty rather than one to an owner. Consider Companions provided to long term prisoners for their mental health: long years of conversation might well inadvertently reveal criminal associates' names, or even confessions, to the authorities.

Safeguards for The Information Content of a Companion

A Companion that had learned intimate details of a user's life over months or years would certainly have contents needing protection, and many interests-----commercial, security, governmental, research---might well want access to it, or even to those of all the Companions in a society. If a society moves to a legal state where one's personal data is one's own, with the owner or originator having rights over sale and distribution of their data (which is not at all the case at the moment in most countries) then the issue of the personal data elicited by a Companion would automatically be covered.

If we turn to the issue of one's data after death, there are already four major types of death site on the internet:

- memorial and tribute sites created for the already dead;
- "locked boxes" of assets and secrets for survivors that protect the individual's interests after death;
- "legacy" sites containing last wishes and emails to be revealed or sent after an individual's death;
- "life story" sites that manage autobiographical material for an individual creator, to leave some form of self-presentation of their life.

A Companion would clearly be useful to the police when wanting to know as much as possible about a murder suspect, so that it might then be an issue of whether talking to one's Companion constituted any kind of self-incrimination, and in some countries that form of communication might in future be protected. Some might want one's relationship to a Companion put on some basis such as that of a relationship to a priest or doctor, or even to a spouse, who cannot always be forced to give evidence in common-law countries.

More realistically, a user might well want to protect parts of their Companion's information---- to have discretion, in the term I used above----- from particular individuals: e.g. "this must never be told to my children, even when I am gone". It is not hard to imagine a Companion deciding whom to divulge certain things to, selecting between classes of offspring, relations, friends, colleagues etc. There will almost certainly need to be a new set of laws covering the ownership, inheritance and destruction of Companion-objects in the future. A Companion would probably be designed as a gleaner of data from open sites such as Facebook and from the user's own holdings of documents and images. But, since some Companion applications will be health-related, there is no reason why this user-based data hoard should not ultimately incorporate health information. In countries such as Spain where everyone has access to and ownership of their own health information one would expect that. This is the opposite way around from the focus of much current public discussion, particularly in the US and UK, where public and private organizations are almost certainly more intrusive than elsewhere, and where as a consequence much public opinion is either negative or ambivalent on these issues.

There is at present a contradiction between the forces of two parliamentary acts in the UK. On the one hand, one is not allowed to publish or even store the names of, say, one's cricket club on the web or a computer without registering one's intention. Nor can one legally publish on a noticeboard the names and marks of one's students, and one may not even be able to publish a remark about someone's health, lest that imply access to privileged health information. There is an obvious contradiction between those constraints and the freedom of access that companies and the State actually have to electronic transactions of all kinds. It is common knowledge that British and American state security process a substantial percentage of our emails and phone records as and when they choose and without any need of access to the courts, even though a journalist has no such protected privilege. As a recent editorial in *Vanity Fair* put it succinctly: governments seem to want "Transparency for you, but privacy for me".

Identity and Identity Shielding

If a Companion is to be an interface to the Internet, say, for a user who is technologically incompetent yet who must conform to the standards of identity that society requires and imposes, then that Companion will have to understand identity to some degree and possibly be able to manipulate slightly different forms of it. In the US and UK, identity is currently established by a range of items with numbers, from passports through credit cards to health, driving licence and tax numbers (some with associated passwords or PINs), with the Social Security number having a definite primacy in the US. In most EU countries there is a single ID number, of which the simplest is on the life-long single *Personnummer* in Sweden. States prefer a citizen to be identified by a single number and in the UK there is currently strong pressure for something closer to the Swedish model, although UK law has, at the moment, no clear definition of identity (with legally registered unique names and addresses, as in most of the EU). There is no legal problem in the UK with having a number of identities simultaneously and bank accounts for each (as a famous case brought, and lost, by the Post Office showed some years ago) so long as there is no intention to defraud.

All this is important since identity checks are the basis of all Web transactions and if a Companion is to deal with a person's affairs it will need something approaching a "power of attorney" or at least an understanding of how identity is established in Web transactions, as well as a method for establishing that its owner approves of what it is doing in individual transactions, in case of later disputes (e.g. by angry relatives after an old person's money has been spent).

A more liberal data regime would be based on *identity shielding,* one where, if our data is open to the state and to corporations, so should it be for individuals to use as they chose. The key principle must be an individual's control of their data for their own purposes, such as their right to know what is held on them, way beyond current minimal access that exists to credit agencies and their records. Dave Birch has expressed a clear view on how an individual could combat state and corporate identification of individuals, by a system of controlled and limited revelation of identity in transactions.

The difficulties with the general principle of "identification minimization" are in the details and, in particular, how we minimize the inevitable loss of information for criminal detection and national security. There are many

technical devices that could be more widely used than they are, including ways to shield our computers from data gathering and linking of data.

Almost certainly, there will be increasing commercial provision of methods of avoiding the identification of one's machine's address, geolocation etc. etc. This used to be associated with anonymous remailing sites and the anonymity of activities such as paedophile rings and terrorist groups. But such facilities will become increasingly available, and Lance Cottrell of NTrepid has argued that disguising one's identity by such methods so as to operate "anonymously" on the Internet will be a growth industry in the coming years, for all the reasons given here. Surveys show that children may already be adapting to this possibility by creating alternative identities on social media. Facebook has recently decided to enforce a, previously little used, rule of "no anonymity on Facebook" even though there has been no way, in general, of detecting such anonymous membership. It is probably revealing that it has raised this issue in connection with its own commercial functioning in China whose government is keen to block anonymous criticism and wants to be able to root out such anonymous critics with Facebook's help.

A simple identity limitation could be used on the UK's DVLA site: the national site that deals with car and driving licences. It checks insurance and car test details and issues a record that it stores and which police and insurance companies can access. But there is no reason why one should have to reveal the name of a car's owner in that process, any more than with one's use of an anonymous prepaid phone. The police have many ways of finding out who is driving a car on a given occasion when they need that fact, and ownership is no guide to that for any given accident or other event; cars are full of biomarkers which could well be sufficient for investigative purposes.

Web "Anonymity" Over-Emphasizes the Importance of Names

Tim Berners-Lee said recently that soon anonymity in computer transactions will be impossible because of the ancillary information available—based on location, past search actions etc.--- from which our identity can be inferred even if we are able to shield our name explicitly. A key necessity for avoiding this (should we wish to) will be some of the mechanisms described above, such as payment mechanisms that are web-usable but as anonymous and as reliable as cash. Bitcoin was designed to offer this but is enormously complex and expensive in terms of computation: the wider use of payments through anonymous digital phones would be an obvious and cheaper start. There are many technical details and possibilities, which could also involve the use of trusted agents, such as Companions, to act for us at commercial and government portals: trusted both by the user and the provider. What is needed is the will to move to such a system if enough of the population find the current situation (and its likely developments) intolerable.

However, in all this one must not be led into overemphasizing the importance of locating a unique *name* so as to identify an individual. The very word "anonymity"--- having no name—leads us into this trap, but Berners-Lee's remark above does not emphasize the issue of names, and this is a crucial point, one which serves to mitigate some of what has been argued for so far. A citizenship roster for the citizens of a state, each having an identification—a personal number, say—is not the same as a link to unique name nor should it be. The French state has traditionally demanded the address of every citizen and that their name be drawn from a roster of acceptable names, but English-speaking societies manage without either of those conventions. One's name is not crucial to any of this identification, since only bio-markers are fully reliable for that, especially since very few names are "google-unique" and their degree of ambiguity varies from society to society depending on social/historical factors: Sweden and Arabic-speaking countries, for example, have relatively small name lists and high name ambiguity. The resolution of such ambiguity is now an NLP technology for settling issues such as "Which George Bush is this document about?".
Names, addresses, and dates-or-birth are only some among many features of use in locating an individual---down to an index number, to a current spatial location, or to a unique biomarker—they are not special in the way we sometimes feel. Having a (correct) name may give us no special "hold" on an individual if we have enough other features, as is known to students of films and television with plots from *Mad Men* to *The Return of Martin Guerre*, where the message is "this is you no matter what you call yourself". Everything short of that is the same situation we encountered when discussing meaning in the Semantic Web: symbols like URL only pan out in the end in more symbols, they never ground in actual hard bound books.

Perhaps the truth here will be that Internet tracking and features give a "cloud" of probabilities as to who an individual really is, and this is sufficient for many commercial (if not all security) purposes: in this cloud the individual's name is only one feature and perhaps not the most important one. This may seem an unduly technical analysis of the issues, but I believe it remains to be shown that there is any distinctively ethical component to the seeking, finding, giving or withholding of an identity, at least in the sense of a name, as seems to have been generally assumed in recent cases such as that of the undercover policemen who conducted affairs under assumed names with those they were spying on.

However, "ownership" of one's identity, one's personal data etc. clearly has an ethical dimension, at least for those who believe that one's ethical sense is tied to one's sense of identity and that it may not be possible to have one without the other. I would prefer to see the issue in terms of providing legal incentives for Internet Service Providers to give identity protection to those who want it, such as by the badging of sites that *require less identification* in order to carry out transactions.

Tim Berners-Lee recently expressed his disillusion with the hate and disinformation-laden web that his creation had turned into, and gave suggestions for personal control of data of the sort I have been discussing. These he presented under the name of a new project called SOLID. It may be that human nature is such that the distancing from face to face communication that the web provides unleashes emotions and incontinence such that no technical constraints can ever remedy and humans will simply have to learn to change their instinctive behaviour.