*Webology, Volume 3, Number 2, June, 2006*



# Chemoinformatics and the World Wide Web: An Interview with Professor Peter Willett

**Yazdan Mansourian**
Department of Information Studies, University of Sheffield, UK.
E-mail: y.mansourian@sheffield.ac.uk

*Received April 5, 2006; Accepted June 1, 2006*

---

## Introduction

Information science is an interdisciplinary area of study which has strong links with a wide range of subjects. As a consequence of these interactions, a couple of branches have been emerged in information science during the past few decades. For example, development of bioinformatics is the result of collaborative research on common grounds between biologists and information science experts. Similarly, Chemoinformatics is an area of research and practice which has roots in chemistry and information science.

This report is the direct transcription of an interview with Professor Peter Willett in March 7, 2006 and seeks to illustrate some general issues in chemoinformatics and its linkage with the related areas. Professor Willett is the head of Department of Information Studies at the University of Sheffield and is one of the well-known researchers in information retrieval (IR) in general and chemoinformatics in particular.

The interview consists of two parts. The first part is related to Professor Willett's specific research area and its relation to the World Wide Web. In the second part he addresses some general issues about his personal experience of using the Web as an end-user rather than an expert in IR.

In the following text PW stands for the interviewee; Professor Peter Willett; and YM for Yazdan Mansourian who is the interviewer.

## PART I: HOW CHEMOINFORMATICS AND THE WEB ARE LINKED?

YM: As far as I know your main research area is Information Retrieval in general and chemoinformatics in particular and the first part of the interview aims to seek your opinion on the possible links between the Web environment and your particular research area. In another word, the first part focuses on your ideas about how the Web affects chemoinformatics and how chemoinformatics have an effect on the Web. However, before beginning the main discussion; for the first question; may I ask how would you define chemoinformatics for somebody who does not know much about it?

PW: Computer techniques that are used for the representation, searching and processing of information pertaining to the chemical structures of molecules. In part, this involves the processing of textual documents as in conventional information retrieval, but of far more importance in chemoinformatics is the processing of the two-dimensional (2D) and three-dimensional (3D) representations of the structures of chemical molecules.

YM: There are a few similar terms including chemiinformatics, chemical informatics, molecular informatics or chemobioinformatics that are used interchangeably; would you consider them as synonym or are these terms different?

PW: I would regard all of these as being effectively synonymous, although if one wished to be very precise then one might argue that, for example, chemical information science was not quite the same as chemoinformatics. There is one related field that I think is distinct: this is chemometrics, which is to do with the statistical analysis of chemical experiments (but which clearly sounds very similar to chemoinformatics).

YM: Which one does describe your research area most appropriately?

PW: I think that chemoinformatics is now the most widely used of these terms.

YM: I presume the history of chemoinformatics comes back to the time before the Web. However, after emergence of the Web all areas of IR have been revolutionized by this new media in many ways and it seems that chemoinformatics should not be an exception here. Nevertheless, chemoinformatics is a very specialized area of study and the Web is a very public media. In your opinion how these two different components have been interacting with each other over the past fifteen years (after the emergence of the Web)? I mean how do you perceive the links between these two issues?

PW: I am going to say something that I expect some of my fellow researchers would disagree with, in that I do not think that the emergence of the Web has made very much difference to chemoinformatics, at least as I study it. There are two

reasons for this comment. First, our interests here in Sheffield - and I say "our" since the Department of Information Studies has been working in this field for some forty years now (even though the word chemoinformatics did not exist until quite recently) - have always focused on the basic algorithms and data structures needed for processing chemical-structure information in various ways, not on the distribution of that information, which is one of the Web's great strengths. Second, much of the basic research, and certainly the great bulk of the applications of chemoinformatics, is carried out in corporate, rather than academic, environments. The Web is a wonderful way of providing of 24/7 access to huge amounts of public data but that is not relevant in a corporate context where chemical structures provide one of an organisation's principal sources of intellectual property and potential profits: this sort of information will not be made widely available on the Web for the foreseeable future. Intranets within companies are incredibly important as a way of facilitating local access to corporate information but not the Internet. Now there are chemical datasets that are available on the Web, most obviously those produced by the *National Cancer Institute* in the USA and those that will be come available as a result of the emerging *PubChem* initiative. But the amounts of data publicly available are miniscule when compared to bioinformatics, where there are vast data repositories that have emerged from academic research and that are publicly available via the Web.

YM: But the Web has been useful even in corporate environment in many ways for example in terms of cooperation between companies or whatever?

PW: It is obviously useful to people within the companies insofar as their staff can access the voluminous amounts of chemical information that are available in the published chemical, biological and medical literatures. But this situation is no different from any other research-based discipline or profession. As to collaboration between companies that happens very little given the intensely competitive nature of the industry.

YM: I think the Department of Information Studies at the University of Sheffield is one of a few institutions in the UK or possibly around the world which offers MSc and PhD in Chemoinformatics, in your opinion how these courses map into the big picture of information retrieval and information science in general?

PW: Well, people still debate as to exactly what information science is but I think it's reasonable to assume that chemoinformatics falls within the general area of information science, in that it involves the processing of information, albeit a very specific type using very specific processing methods. In just the same way, here in our department we have courses and modules that focus not just on chemoinformatics, but also on health informatics and educational informatics; and there are many other specialised informatics courses in other institutions that also surely come under the broad heading of information science. I think that the part of information science that is most closely related to chemoinformatics is information

retrieval (or IR): in just the same way as IR researchers develop novel ways of processing textual information (and, more recently, multimedia information) so researchers in chemoinformatics develop novel ways of processing chemical-structure information. Indeed, it's my belief that there is sufficient commonality to enable some techniques that were developed in the textual domain to be applied in the chemical domain and vice versa. Here in this department, I'd like to think we have been quite good at cross fertilization over the year. It is for this reason that, even though I stopped working in the textual IR area a decade or more ago, I still try to keep an eye on what is happening there to see if there is an idea that might be applicable to the chemical domain. So, whilst I don't read every issue of the main IR journals in the way that I would have ten years ago, I do try to skim through the contents pages of these journals and of the annual *SIGIR conference*, so I am aware of what is going on.

YM: The reason that I asked this question was because I was wondering in your opinion whether chemoinformatics is closer to information science or to chemistry?

PW: I would say the chemoinformatics is closer to chemistry than to information science. That is said, Chemoinformatics covers quite a wide range of things and certainly here in Sheffield we have focused upon those aspects that are related to algorithms and data structures, rather than upon the actual application, so our studies are more at the information science end-of-things. As an example, there is a technique called *docking*, that essentially finds molecules that have a particular shape. In collaboration with industrial partners some years back we developed a program, called *GOLD* that is now one of the most successful commercial programmes for *docking*. Lots of people in industry use this programme to carry out sophisticated analyses of how molecules interact with proteins: we don't have the chemical and biological knowledge to carry out this sort of analysis, but we can provide the underlying computational tools. In much the same way there is a technique called *QSAR*, which stands for quantitative structure activity relationships and which involves developing mathematical models that relate the structure of a compound to its biological properties. We have been involved in the development of new tools for this, but it is for practicing medicinal chemists to understand and exploit the information that you can get from the applications of the tools.

YM: What is your general advice to potential students who might be interested to continue their education in chemoinformatics? I mean what would you consider as the pivotal feature/qualification of a young researcher in chemoinformatics?

PW: If someone has that interest then I would say send an application to the Department of Information Studies at the University of Sheffield or one of the limited number of other institutions worldwide which offer such dedicated courses, nearly all in chemistry rather than information-science departments. An MSc course would be a good starting-point I think doing our courses or others in master

level would be a good starting point. If they want to go on to work in industry then it would be worthwhile obtaining a PhD, as these are valued very highly in the pharmaceutical and agrochemical industries (the main users of chemoinformatics systems and industries that live or die by the quality of the research that is carried out). Given the limited numbers of people with knowledge of and qualifications in chemoinformatics, having either an MSc or a PhD in the subject means that you will find it very easy to get a good job.

YM: One of your research areas is "citation-based analysis of research performance" is this issue related to Scientometrics?

PW: Yes, scientometrics, bibliometrics, and citation analysis all cover very similar, if not identical, areas. It is not an important interest of mine but over the years I have published a few papers on it. I certainly wouldn't go as far as some of accepting some of the claims that are made for citation analysis; that said, it has certainly been my experience that most forms of citation counting yield results that are in line with one's gut feelings and that measure something sensible. Thus, while I wouldn't go along with the idea that you can do away with the entire Research Assessment Exercise simply by counting citations, I can't see any reason why citation count shouldn't be a performance indicator that people can use.

YM: So is it a good way to measure the performance?

PW: It is a very simple thing to measure and, more importantly, the numbers that come out (either from the conventional citation indices or from Google Scholar) correlate with other performance measures based on peer review. Thus, the numbers encode meaningful information.

YM: In "citation-based analysis of research performance" what do you mean by research performance?

PW: Research performance is the quality of what is carried out. There is the problem that quality is one of those words like "excellent" that have become hackneyed with repetition, but essentially it means how good is a piece of research. Now, one can normally only assess quality some years after the work was done because only then can you see whether it had any effects. The great bulk of research either has no effect or becomes assimilated without people being specifically conscious of it; but there are papers that do get cited and often over quite long periods. I know people cite for all sort of reasons but by and large people cite because they have read something that is meaningful, important and useful for what they are doing: whilst I agree there are lots of exceptions, I think it is very difficult to disagree with the view that something that is subsequently cited by a large number of people made some sort of contribution to the development of a subject. Moreover, if something is cited several hundred times in the literature then that is more likely to be a more significant contribution than something that has

been cited infrequently. I would argue with the view that something cited 252 times, say, is more important than something cited 227 times but not with the view that the former is more important than something cited 5 times.

YM: One of the areas in bibliometrics is *Webometrics*, what is you general idea about the current trends and issues in *Webometrics*?

PW: I wouldn't feel confident to make an answer. I only published one paper on it several years back, based on an MSc dissertation, when *Webometrics* was a very new idea. I have not done anything since then: go and look at the work of people like Mike Thelwall, who have built up a very impressive body of research.

YM: As far as I know, the research that you are involved mainly relates to technical aspects of information storage and retrieval (e.g. computational tools for molecular diversity analysis; similarity searching in databases of 3D molecules and macromolecules and so on), in your opinion how important is to carry out some user-oriented studies in Chemoinformatics? I mean how much Chemoinformatics researchers know about the perceptions, feelings and behaviour of end-users of their products? I mean who are end-users in Chemoinformatics and how much knowledge does exist about this group's information seeking behaviour and their information needs?

PW: The short answer is they know a huge amount. Chemoinformatics is a rather strange area since the research has been done at least as much within companies as it has in academe, and if the researchers within the companies were not closely attuned to the need of their users who are in the lab next door, then they will be out of a job - so there is a very close interaction! To repeat what I said earlier our particular interests here in Sheffield are rather more in the algorithmic techniques that can provide the foundations for system rather in building systems themselves: hence as long as the technique can do something useful then it is for others to customize it and make sure it has all the bells and whistles need to make it highly usable. However, before you get the idea that we are typical ivory tower academics who don't care about the real world, the majority of our research is carried out in collaboration with, or funded by, pharmaceutical and agrochemical companies....and the reason that they have make the investment is that we develop techniques that they can then implement for their chemists and biologists. Thus, usability and usefulness is a very important part of our chemoinformatics research, but it is not the bit that we ourselves do.

## PART II: PERCEPTION ABOUT THE WEB AS AN END-USER

YM: What is your general feeling about searching the Web as a user and not as an expert in IR?

PW: I am an academic who has been working in a very specialised field for a long period: thus unlike many, or possibly most, people who use the Web I tend to know a huge amount about what I am looking for. Thus, rather than just logging on for a general wander around looking for things, I normally go looking for a specific piece of information, and I guess that might well be true for many professionals. So I might be looking for a specific thing like a phone number or a pharmaceutical company's last annual report or somebody's address. Very little surfing and thus my experience of searching the Web may not be typical of many users.

YM: How long have you been searching the Web?

PW: Well since Mosaic in 1994.

YM: How often do you search the Web?

PW: My browser is switched on with my email when I come into work at about ten to eight every morning and it is switched off when I go home about quarter past five, so it is running all days long. I have never counted how often I use it but I do have three browsers always open - one has the university internal phone directory open, one has the university library A to Z list of online journals, and the third is used for searching.

YM: How satisfactory are search results for you in general?

PW: Because I am looking for precise information, it is normally very high. I wouldn't give a figure other than noting that one search failed today because I got a page 404 "*the page not found*" response - but that is the fault of the website and not my fault.

YM: Do you always manage to find what you want on the Web or do you ever not find what you have been looking for?

PW: You can't always manage, and end up swearing at it and banging the screen like everybody else....but that is pretty rare.

YM: Do you ever have the feeling that there should be more relevant information about your search topic on the Web but you are not getting to it?

PW: Generally I'm looking for a fact and once found that is it. It is not like doing a subject search when you start a PhD and you want to find all the documents about your subject that have ever been written. YM: In your opinion, how likely is it that you have missed something about your search topic even if it is specific information?

PW: I'm not perfect so I am sure there are times that I fail: you recall that you say a website once but cannot track it down when you actually need it....but there's

probably a colleague down the corridor or an email from whom you can get what you need.

YM: How much does it matter to you if you know you have missed something while searching the Web?

PW: I am not sure how you can answer a question like that. As an extreme example of something mattering, let's assume I am a patent officer for a company that wants to patent a new chemical compound as a potential drug: then it is absolutely vital that I be certain that I haven't missed anything. There are few things I do that would demand that level of certainty.

YM: Could you describe how "*The Invisible Web*" might mean to you?

PW: as I understand it, the *The Invisible Web* refers to those places that are not visited by crawlers, and that you are hence unable to access. I guess that the size of the *The Invisible Web* would be inversely related to the number of different search engines that you use, so I guess the size of the *The Invisible Web* would be rather smaller if you used multiple search engines rather than just Google, or whatever.

YM: Is there anything else that you would like to add?

PW: No that is okay.

YM: Thank you indeed for your time.

PW: You're welcome.

---

***Bibliographic information of this paper for citing:***

Mansourian, Y. (2006). "Chemoinformatics and the World Wide Web: An Interview with Professor Peter Willett." *Webology*, **3**(2), Article 27. Available at: http://www.webology.ir/2006/v3n2/a27.html

---



---