From *The English Poetry Full-Text Database* to seven flavours of *Literature Online*: ten years of digital publishing in the humanities at Chadwyck-Healey, 1991-2001, and a look into the next ten.

[1]

When I was invited to speak at this conference it seemed like an ideal opportunity to take a retrospective look at Chadwyck-Healey's digital publishing in the humanities. It's ten years since we began work on digitising the humanities, and I think it's a reasonable claim that during those last ten years we have been the world's leading commercial publisher in the field. We aim to stay that way, continuing our investment in the digitisation of the primary works of literature and history, and in an increasing number of secondary works as well – reference, criticism, bibliography, biography and the like.

It also seems especially appropriate to give this paper here in Sydney, as it was a library in Sydney – the State Library of New South Wales – which placed the very first order in April 1991 for our first full-text database in the humanities, *The English Poetry Full-Text Database*; and it was a consortium of 16 Australian libraries, including the University of Sydney, and chaired by the then University of Sydney Librarian, which later in 1991 made a joint commitment to acquiring *English Poetry*, a commitment which gave us tremendous encouragement at a time when the project was proving editorially, technically and financially challenging. Without that early expression of faith in our publishing vision, and our ability to deliver it, who knows whether we would have gone on to achieve as much as we have done in these last ten years. So, thank you Australian librarians and academics.

Since then we have published almost 50 discrete databases in the fields of literature and history, comprising essentially historical materials which we have digitised from printed sources. **[2]** These include 20 different full-text databases in English and American literature, plus the very large database of the *Annual Bibliography of English Language and Literature*, which we digitised from its first volume in 1920; **[3]** 7 full-text databases in German literature; the two very large Latin databases, the *Patrologia Latina* and *Acta Sanctorum*, and two French and one Spanish database; [**4**] and around 10 databases in history, primarily indexes such as the two great indexes to *The Times* and the complete index to the British *House of Commons Parliamentary Papers* from 1801 to today, and *The Times* newspaper in image form from its first issue to 1870.

[5]

This year alone, we have been working on eight full-text databases in parallel: four in English and American literature – the second edition of *English Poetry*, which includes a substantial number of Australian and New Zealand poets, new releases of *Twentieth-Century English Poetry*, *Twentieth-Century English Drama* and *American Drama*; the complete edition of Luther's Works, which we will complete in 2002; the first releases of the *Digitale Bibliothek Deutscher Klassiker*, the electronic edition of the great German series published by Insel Verlag, the *Deutscher Klassiker Verlag*; *Acta Sanctorum*, which we will also complete next year; and the *Annual Register*, the great historical review which we are digitising back to its first volume in 1758, and which we will complete this month – a work which will benefit enormously from being databased. This already large list of digital publishing projects also omits the many works we are digitising and adding to our online services but which are not

published in their own right as individual databases, such as individual works of literature to supplement an existing collection, or an individual reference work. It also omits *Periodicals Contents Index – PCI* – for which we have now digitised more than 12 million article records from more than 3,000 journals, and have scanned complete runs of almost 150 journals.

[6] These are just some of the statistics of our investment in digital publishing in the humanities over the last ten years or so. The investment of Aus$60 million is probably at the low end of estimates, and it represents only the investment in in-house editorial work on the texts we have digitised and the outsourced keying and scanning; not the investment in the scoping of projects, the fees to editorial advisers, the selection of content, its sourcing, the software development and database building, the storage and delivery of the databases, and the royalties to content providers.

But it's not these individual databases which I want to focus on today, **[7]** but the online services in which they are included, and the transition from our publishing of individual databases, primarily on CD-ROM and magnetic tape, to the publishing of far larger and more complex online databases, or services.

I had considered calling this paper 'from Digitisation to Publishing', reflecting the much greater use of a broad range of publishing skills as we moved from a publishing model led by CD-ROM databases to one led by online services, but that would be unfair on my Publishing colleagues who have worked on these individual databases. There was very significant editorial investment in *The English Poetry Full-Text Database*, for example, in the selection of authors, works and editions; there was

significant editorial investment in the selection of the 21 Bibles in *The Bible in English*. Even in a work like the *Patrologia Latina*, where the only issue of selection, once we had decided to do the project, was which edition of the *Patrologia* to use (and that wasn't, in itself, a difficult decision), even in this database there was a very significant editorial investment in the coding scheme, in our use of SGML to encode and make separately identifiable and searchable the different components of the text, for example coding Biblical citations as such, to enable researchers to retrieve those elements of text from all other elements. And of course we have made a similar investments in the encoding of all our digital databases, the level of investment depending on the complexity of the source work. This probably reached its greatest height, or depth, in the Weimar edition of Goethe's Works, with its very extensive and complex critical apparatus, in which the coding scheme enables the researcher, should he wish to identify, for example, those passages of text which Goethe annotated in his manuscripts in red ink, say, as opposed to black pencil.

So in each of these databases there was a substantial investment in editorial processes, in the value which publishing can add to a work. But it would also be true to say that by far the greatest share of the investment, possibly as much as 90% of the cost, was in the digitisation process: in the acquisition of the source texts, their copying, their mark-up according to the schemes that our editors had devised and, above all, in their double-keying and proofing; and then in the combination of the digital texts with software to create products which researchers could use.

And this is how it was for the first five years or so of our digital publishing in the humanities, from 1991 when we began work on both *English Poetry* and the

Patrologia Latina, to 1996, when we began work on *Literature Online*, which we launched in December 1996.

Now the investment in *Literature Online* wasn't significantly different in kind to the investment in individual databases. The production processes were different, but the investment was still primarily in the data and in the software required to access the data, and in the interface design. In moving from CD-ROM to the Web, we moved from one software platform to another, and this involved some work on the data, as did the requirement in *Literature Online* for the user to be able to search across multiple databases. We had some substantial work in standardising the SGML coding schemes to enable this to happen. At launch we brought seven full-text databases together and structured the service so that users could search in similar ways within individual databases as they had been able to do on CD-ROM; and so that they could search in somewhat less sophisticated ways across all seven databases, in poetry, drama and prose. We also included in the service a bibliography – the *Annual Bibliography of English Language and Literature*, or ABELL, an English dictionary, the *King James Bible* and links to relevant external websites.

I will not play down the effort required to move these seven databases from CD-ROM to the Web. The project – which I led – took 11 months of work from a large team of programmers, data conversion specialists, interface designers, editors and marketeers. It required a different production process from that for the CD-ROMs; for the CD-ROM databases the process was essentially serial – we identified the texts, sourced them, copied them, coded them, sent them away for keying, got them back and proofed them, both texts and coding, parsed them, built the database and combined it

with the search software and the CD-ROM interface; for the development of *Literature Online*, most of the production processes happened in parallel, which meant that management and coordination of the project was significantly more challenging. And we had to do some entirely new things, such as creating what we called the 'Master Index' to the site which enabled a user to find authors and works in *Literature Online*, regardless of which individual database they were in, or of whether they were on an external website to which we linked. Nowadays, of course, the Master Index would be called metadata, but that's the name we gave it at the time.

There were other important changes in the move from CD-ROM to the Web, most notably the move from selling databases in physical form on perpetual licences to selling a collection of databases online on an annual subscription basis, if you like, the move from selling products to services, a cultural and commercial change whose full implications we only partially understood at the time. But these issues have been explored in other papers and I'd like to focus in the rest of this presentation on a different aspect of the move from CD-ROM to the Web, an aspect which has only truly developed in our digital publishing in the last two to three years.

In spite of all the differences between the individual CD-ROM databases and *Literature Online*, both served primarily the same user: the university academic or postgraduate researcher. They were essentially research tools for the advanced researcher. A scholar could use the *English Poetry* database, or all the databases within *Literature Online*, to search for the first occurrence of a poetic phrase, to look for similarities between the works of one writer and another, to look at the influence of the Geneva Bible on Shakespeare's language, and the influence of successive

editions of the Bible on the development of the English language overall, or just to check a rare text not otherwise available in his or her library. But for the first- or second-year undergraduate, or the lay user of a public library with an interest in English literature, neither the individual CD-ROM databases nor *Literature Online* truly met their needs. All that raw primary text, with no critical apparatus, with no biographical information, with no placement of the writer in his period, unmediated by any editor, the overall lack of context, made them difficult tools for the student to use. For the researcher, the ability to search the texts and to identify relevant journal articles and other criticism from the 70+ years of the ABELL bibliography, were a boon; for the student, the databases were probably somewhat daunting.

So in the last three years or so it is in this area, in the value of *Literature Online* to the less experienced researcher, and to the teacher and learner, that we have made the most significant changes. And in doing so, we have gone from Digitisation to Publishing, in the sense that our editorial investment in these services has begun to match, and in some cases to exceed, our investment in pure digitisation. And this has been a very exciting development for us. We are commissioning material of our own, for example biographies of writers and extensive captions for illustrations, photographs and other image material. We are investing heavily in the organisation and structuring of the digital content and in the provision of access to it through topic trees. We are working closely with teachers to ensure that our online services meet their teaching needs and the needs of their students. While we continue to digitise on a large scale, it is in the context of publishing which is driven by what our different groups of users tell us they need.

There are now seven versions or flavours of Literature Online and I'm going to focus on three of them today to illustrate the development of our digital publishing in the humanities: *Literature Online Complete*, *Literature Online Select* and *ProQuest Learning Literature* – a different brand but produced by Chadwyck-Healey and based on the same data warehouse of primary and secondary content.

[online demo]

[8] *Literature Online Complete* is the complete collection of 330,000 primary works of poetry, drama and prose, along with extensive collections of reference materials such as the *New Princeton Encyclopedia of Poetry and Poetics*, the *Columbia Dictionary of Modern Literary and Cultural Criticism* and the *Encyclopedia of Post-Colonial Literatures in English*; the complete ABELL bibliography, now including material not in the printed edition; the full text of 65 literature journals; newly commissioned biographies of 350 writers; bibliographies for 1,200 writers; and links between the texts of works in *Literature Online* and the images of those works in *Early English Books Online*, for mutual subscribers to the two services.

Literature Online Complete is intended to support postgraduate research in English and American literature and more advanced undergraduate study. It is designed for the larger universities predominantly in the English-speaking world with large departments of English. In our design we have attempted to retain all the features that the subject expert requires, while making it easier for undergraduates with a reasonable knowledge of the subject to get the most from it too, and to undertake their own research.

So if the researcher wants to find references in poetry to the 'deep blue sea', then he can use the 'Search Texts' option to do so, and get a list of relevant works, and view them on screen **[9-11]**; while if the undergraduate wants to find out what *Literature Online* has on Chaucer, then he can use 'Find Authors' to do so **[12]**, and will be given the option of reading a biography **[13-14]**; going to the works themselves **[15]**; reading about the author and his works **[16]**, in journal literature, in reference works and elsewhere on the Web; or going to a bibliography of his works, regardless of whether they are available in *Literature Online* or not. He can also limit his search just to Criticism and Reference **[17-18]**. Everything is organised in one place but the user still has considerable freedom to select the materials he wishes to make use of.

[19] As you can see from the home page, *Literature Online Select* has a very different feel to the *Complete* database. It contains a subset of the primary texts – around 200,000, from 1,200 authors, with the emphasis on the most widely studied authors, especially twentieth-century authors. It then contains around 8,000 selected secondary works – journal articles and other criticism - plus biographies of around 900 authors, a reference shelf of reference works and links to around 1,000 websites. But it's not only the contents of the database which differ, but its organisation. **[20-21]** While you can search the texts as you can in *Complete*, with most of the functionality, the primary point of access is the Study Pages **[22]**, around 540 pages covering authors and topics **[23]** which bring the biographical information, the works, the criticism and the links to web sites together in an easy-to-use way. You will also see that the biography **[24]** has been edited for this audience, is somewhat shorter and is organised into topics.

Literature Online Select is intended for the undergraduate institution which focusses on teaching bit which wishes to provide its students with a broad resource in which they can do their own research. It still provides access to a wealth of primary material, possibly well in excess of what the institution's library might offer, but it also organises much of that content in an easily navigable way for the learner.

[25] We then move on to *ProQuest Learning Literature*. This started life as *Literature Online for Schools* but has been rebranded as the first of a series of online services for the 11-18-year-old school student **[26]**. We have since launched *ProQuest Learning History* and further subjects will be covered in the next year.

Here you see a different look-and-feel still. There is a section for 16-19 year-old students which is not that dissimilar to *Select* **[27-29]** but when we get to the section for 11-16 year-olds a rather different set of functionality. You can still search the poetry texts, but in a more simplified way, though we offer the extra functionality of a search by topic **[30-32]**. And we also have the Study Pages **[33]**, where you will find a picture of the poet, a short biography and the works themselves, in some cases with glosses and with audio **[34-35]**. There are also picture galleries **[36-37]** with illustrations with extensive captions, commissioned by us, which help to provide a stronger context still for the works themselves.

ProQuest Learning Literature is designed primarily for the UK schools market, with its content reflecting the national curriculum and the requirements of the various examination boards. It also meets the needs of the International Baccalaureate schools. A US version of it is also available.

Across all of these literature services we shall be adding much more multimedia content in the coming year – more audio content, more video content, more images. Indeed, we have just commissioned a series of films of leading poets reading both their own works and those of leading poets of the past which will be added, in one form or another, to all three services that we have just looked at.

As you can see from these slides **[38-41]**, we are now doing the same thing in history, with a range of services to meet the needs of everyone from the school student to the postgraduate.

So I hope that you can see from these short demonstrations how far our digital publishing in the humanities has moved in the last ten years, from the compilation of very large digital collections, published in physical form, for the postgraduate research sector, with the majority of our investment being in the digitisation of the primary texts; to the creation of online services designed to address the needs of different levels of the educational community, from the secondary school student and teacher, through the university undergraduate to the postgraduate researcher, combining even larger collections of primary materials with secondary materials such as reference works and criticism, with a growing body of content commissioned by us and uniquely available to us, delivered online, with a far larger proportion of our investment going into the creation of new content, the selection of content, the shaping of the service and the design of functionalities, navigation paths and interfaces for specific groups of users: in taking a discipline and building services to

meet the needs of users with different levels of expertise in that discipline; enabling all of them to research at their different levels of ability.

We can identify three periods to our digital publishing so far: the first period, of CD-ROM publishing of large, generally finite collections of works of literature or history, from 1991 to 1996 (though we continue to publish such databases for those institutions which prefer to 'own' such databases on a perpetual licence, or which wish to mount them on their own servers); the second period, from 1996 to around 1999, when we moved from one medium to another, built ever larger databases, sold them on a subscription basis rather than outright, but still essentially served the needs of the postgraduate researcher rather than the student or teacher; and this current period, beginning around the end of 1999, in which we have begun to shape services, based on our huge collection of digital content, for the needs of widely differing user groups, and for which we are increasingly creating our own content and investing as much in the 'publishing' of it as in the 'digitisation' of the literary and historical materials. The development of this third stage of our publishing was very significantly accelerated by the acquisition of Chadwyck-Healey by what is now ProQuest Information and Learning in October 1999. While Chadwyck-Healey had already begun the development of a literature service for schools, the access that ProQuest provided to its vast digital store of periodical content, and to its Digital Vault of historical collections such as *Early English Books Online*, has enabled the Chadwyck-Healey publishers to select from a much wider range of content and to create much richer services much more quickly.

So how long will this third stage last, and what will the fourth stage look like?

In preparation for this meeting I looked at a paper I wrote on the development of *Literature Online* in October 1997, in which I tried to predict how it would develop in the next few years. By the time the publisher got round to publishing my paper, a year later, most of my predictions were wrong. At the time of writing the paper we had a Writer-in-Residence in Literature Online, first the Irish poet Matthew Sweeney and then another Irish poet, Eavan Boland. The writers-in-residence commented on their works and invited discussion with users. I predicted that this interactive, dynamic element of the service would grow. There was almost no interest in it and even less use, and we dropped it. We were planning to break the service down into chronological subsets, but again there was little demand for this, beyond the indexing of the content by literary movement and literary period, and we have done this. Some of what I predicted was right, in particular the development of better navigation paths through the content, the richer indexing of the content and the licensing of far more twentieth-century content. In October 1997 we had persuaded poetry publishers such as Carcanet to license their content to us, but we had still to persuade Faber and Faber to do so. In the last four years we have added most of the core corpus of twentieth-century British and American poetry, through agreements with publishers such as Faber and with the literary estates of writers such as Wilfred Owen. But I did not predict the development of a specifically undergraduate service, nor of a separate reference service, nor of a schools publishing programme.

So what can I safely predict for the next few years? **[42]**

The first thing is the continuing digitisation of works of literature, adding ever more to the primary collections. Projects currently underway include additions to the *English Poetry* database, a database of *Canadian Poetry*, the completion of the *American Drama* database and the launch of a database of twentieth-century *English Drama*. And we shall, of course, begin to include twenty-first century works as well, especially poetry. However, the newer databases will not necessarily be published as single large collections at one time, but will be fed into *Literature Online* and its subsets on a regular basis, and then, if there is demand, spun out as discrete collections for those institutions which still want to 'own' them in physical form.

The second is the continuing addition of secondary content – reference works, bibliographies, biographies, journals. In the latest release we have added four new full-text journals to the service, taking the total to 65, and three new reference works.

The third is the continuing addition of content that we have commissioned, including extensive new author biographies by academic experts and in-depth studies of individual works to help undergraduates better understand them. These we call *KnowledgeNotes* and we are about to add around the first 150 of them, and then continue to add them at the rate of ten per month.

The fourth will be the continuing refinement of the structure, functionality and interface design, to ensure that it is as easy to use as possible, regardless of whether you are in the first year of higher education or have been researching English literature for the last forty years, and that it provides value to both types of user.

After that, I think it is down to those of you who use the service to tell us what you want. Having now digitised such a large amount of the primary content, in particular in poetry and drama, we can choose to invest our resources in whatever elements of the service they bring most value to. So it could be works of twentieth-century fiction; or it could be more and more reference and journal content; or it could be an ever greater amount of unique commissioned content to bind the whole thing together and fill gaps; or it could, as now, be some combination of the three.

However we develop this service, and our other electronic services and publications in history, German literature, medieval studies and so on, I am certain that the next ten years of digitising the humanities at Chadwyck-Healey will be as demanding, difficult, frustrating, unpredictable, and as exciting and rewarding, as the last ten.

Thank you for your time and attention. **[43]**

Steven Hall

Senior Vice-President and General Manager

ProQuest Information and Learning

September 2001