LIBER QUARTERLY 18 (3/4), December 2008 – ISSN: 1435-5205. P413-423
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

Improving Access to European E-theses: the DART-Europe Programme

Martin Moyle

Digital Curation Manager, UCL Library Services,
UCL (University College London),
m.moyle@ucl.ac.uk

This paper is based on a presentation given by the author at LIBER's 37th Annual General Conference, Koç University, Istanbul, 2008. Slides available at http://eprints.ucl.ac.uk/9009/ or at http://www.ku.edu.tr/ku/images/LIBER/dart_europe_liber2008_mmoyle.ppt

Abstract

DART-Europe (Digital Access to Research Theses - Europe) is a partnership of research libraries and library consortia who are working together to improve global access to European research theses. The Programme is endorsed by LIBER (Ligue des Bibliothèques Européennes de Recherche) as part of the work of the LIBER Access Division, and it is the European Working Group of the NDLTD (Networked Digital Library of Theses and Dissertations). DART-Europe serves as a European networking forum on issues relating to electronic theses. The DART-Europe partners share an enthusiasm for open access to research theses, and they have helped to provide researchers with the DART-Europe E-theses Portal, a service which enables the discovery of the open access research-level e-theses offered by institutions and consortia from a growing number of European countries. This article gives an overview of DART-Europe, its progress and its future plans, with particular reference to the DART-Europe E-theses Portal.

Key Words: DART-Europe; e-theses; ETDs; Dublin Core; OAI-PMH; metadata harvesting.

Background

DART-Europe began in 2005 as an 18-month project. The founding academic partners were UCL (University College London), Trinity College Dublin, Oxford University, and Dartington College of Arts, working in association with ProQuest, DART-Europe's technology partner during its first phase. The project was resourced entirely through partner contributions, and had several aims, as follows:

- The creation of an open access, free at the point of use portal for research-level European e-theses.
- The creation of a hosted repository service, to support institutions without local repository provision.
- The creation (or collation) of guidelines and the provision of general advice on e-theses management, including issues such as IPR and metadata.
- The demonstration of a digital preservation service for e-theses.
- The identification of a business model for the long-term sustainability of the portal, founded on income from value-added services such as copyright checking and the sale of bound copies.

With hindsight, this was a rather ambitious set of aims, especially for an unfunded and relatively short project! Largely for want of time, not all of these aims were fully achieved. The partners made some progress, however: a portal service was successfully demonstrated, as was a hosted repository service (now available directly through ProQuest). The work on guidance was undertaken to a limited extent, but was superseded by the excellent work of bodies such as GUIDE (Guiding Universities In Doctoral E-theses) and EThOS (Electronic Theses Online System, UK) which emerged in this period. A digital preservation service was not tested, and the partners did not identify a sustainable business model during DART-Europe's first phase.

DART-Europe entered into its second, current phase early in 2007. The partnership is now fully community-led, consisting of research libraries and library consortia. It continues to be resourced solely through partner contributions. It is administered by UCL, and governed by a Board of partners, which is co-chaired by the Library Directors of the University of Nottingham and UCL. DART-Europe is no longer a project, but an ongoing programme, sponsoring

various activities in support of the management, discovery, usability and preservation of e-theses.

The Partnership

At the time of writing, eleven European countries are represented in the DART-Europe partnership, which has the following members:

- BICfB (Bibliothèque interuniversitaire de la Communauté française de Belgique), Belgium
- CBUC (Consorci de Biblioteques Universitàries de Catalunya), Spain
- Deutsche Nationalbibliothek, Germany
- DiVA (Digitala Vetenskapliga Arkivet), Sweden and Norway
- Dublin City University, Ireland
- Ecole polytechnique fédérale de Lausanne, Switzerland
- Helsinki University of Technology, Finland
- Lund University, Sweden
- NORA (Norwegian Open Research Archive), Norway
- Oxford University, UK
- Tartu University, Estonia
- Trinity College Dublin, Ireland
- UCL (University College London), UK
- University of Debrecen University and National Library, Hungary
- University of Nottingham, UK.

The partnership is a mix of national libraries, consortia and research-led institutions, all of whom are contributing in some way to the DART-Europe agenda. Several institutions and consortia have expressed a willingness to join DART-Europe, and it is expected that the number of partners will continue to grow in the coming months.

Organisations wishing to join DART-Europe are invited to sign an Agreement. This is a deliberately 'light-touch' document: it is designed to create a sense of shared purpose rather than to place any significant obligations on the partners. Signatories agree to support seven principles: in short, partners are invited to contribute metadata, on their own terms, to the portal; they are

invited to contribute resources to support DART-Europe's work; and they agree to help DART-Europe to be an effective network for e-thesis information, expertise and resources. The DART-Europe Partnership Agreement[1] is available in full from the DART-Europe website.

Key Relationships

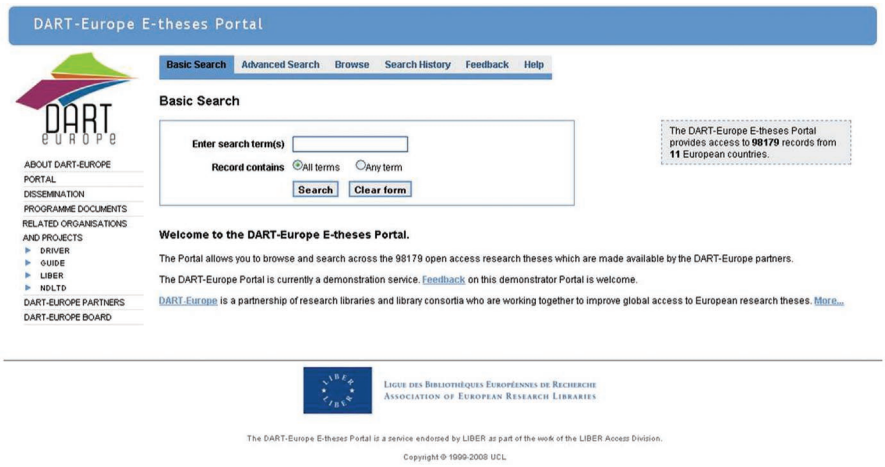
DART-Europe maintains close links with a number of other organisations. The DART Programme now carries the endorsement of LIBER (Ligue des Bibliothèques Européennes de Recherche), and its work has been adopted as part of the work of the LIBER Access Division. LIBER and the Koninklijke Bibliotheek, the National Library of the Netherlands, have signed a Memorandum of Understanding[2] based on a shared vision of perpetual access to digital publications for Europe's libraries and researchers, and DART-Europe is helping to take that Memorandum forward with some investigative collaboration on the digital preservation of e-theses. DART-Europe is also undertaking some exploratory collaborative work with DRIVER (Digital Repository Infrastructure Vision for European Research), with a view to the possible adoption by DART-Europe of DRIVER's infrastructure, for the DART-Europe Portal, while DART-Europe is contributing to a future iteration of the DRIVER metadata guidelines[3] for content providers. Finally, DART-Europe is the European Working Group of the NDLTD (Networked Digital Library of Theses and Dissertations).

The DART-Europe E-theses Portal

DART-Europe maintains a portal for the discovery and retrieval of the open access research theses which are made available by the DART partners. Figure 1 shows the Portal's welcome page. The Portal uses OAI-PMH[4] (the Open Archives Initiative Protocol for Metadata Harvesting) to aggregrate metadata from the e-theses repositories which are maintained by the partners, and offers discovery functions across that metadata, with links through to the full text of theses at the source repository. The Portal is a single entry point for researchers and the public to Europe's doctoral research. It offers

— at the time of writing — access to over 98,000 electronic European theses, with daily updates. Thesis details from over 150 awarding institutions from across 11 European countries are represented in the aggregated content. The DART-Europe E-theses Portal offers benefits across the academic community. Contributing metadata to the Portal is easy for data providers, and they benefit from the increased visibility of their local collections and repositories which it helps to bring about. Thesis authors also benefit from the increased exposure which the Portal adds to their research. Researchers and other users are able to reap the benefits of aggregation — the Portal provides one-stop access to freely available, quality-assured research content from across Europe.

Fig. 1: DART-Europe E-theses Portal: welcome page.



The Portal offers both a simple, 'single-box' search and an advanced search, which allows researchers to specify keywords in the author, title and description fields, and to limit their searches by country, data source, awarding institution, language, and date range. Browsing by date, author, data source, awarding institution and country is also possible. Any selection from a results list may be expanded to show the full metadata record, as harvested by DART-Europe from the source repository; and each full record houses a link through to the full text of the thesis, which is openly and freely available to the researcher. The content of the Portal is also indexed by search engines. This provides an additional point of exposure for the DART-Europe partners'

e-theses, and DART's presence in the wider network environment also offers opportunities for the serendipitous discovery and investigation of the Portal by researchers.

The Portal: Technology and Metadata

Data are collected for the Portal using the open source PKP (Public Knowledge Project) Harvester2 package, and presented through a bespoke user interface, written in PHP. The recent move from the default Harvester2 interface to an in-house development has allowed greater flexibility in customisation for the UCL team which manages the Portal: supporting, for instance, the 'neutralisation' of special characters (so that, for example, a search for either 'Tromso' or 'Tromsø' will produce identical results, regardless of the form in which the word is stored), the easy introduction of new browse indexes, and the agility to deal with data from different consortia, each of which can have different models for data provision and differerent preferences for its exposure.

Metadata about electronic theses is collected in simple Dublin Core (DC) format, although Harvester2 also supports the MODS and MARC standards, and is extensible to support other metadata standards. In principle, the metadata harvesting process is straightforward. An approved thesis is deposited in an institutional, consortial or national repository. Metadata about the thesis is exposed in standard DC format through the repository's OAI interface, which the DART Portal software queries daily for new records.

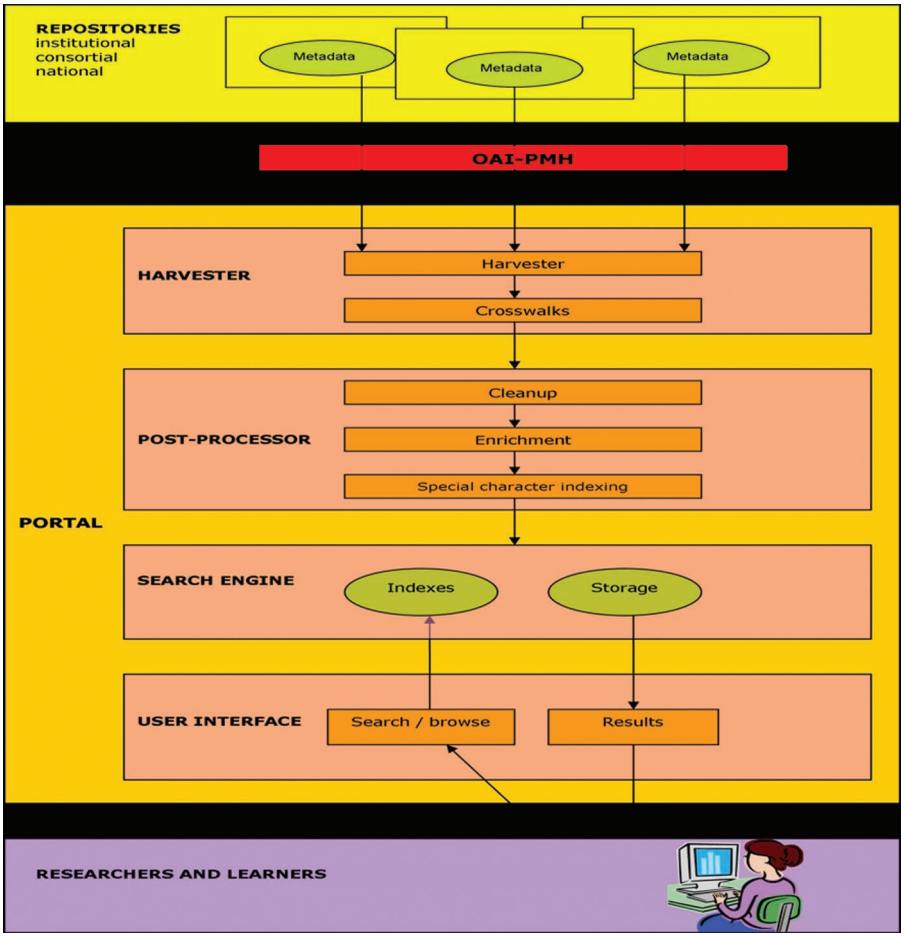
In practice, building services on top of metadata harvested in simple Dublin Core presents several challenges. All the DART-Europe partners have well-organised repositories with rich and meaningful metadata associated with their e-theses. However, there are limitations on the ability of Dublin Core to accommodate such a wealth of detail in a way which is sufficiently well standardised to support third-party services such as DART-Europe. Dublin Core, the Web's *lingua franca*, is first limited by the fact that it forces the reduction of complex descriptions to no more than 15 basic elements. It is simultaneously so versatile that these elements can easily be interpreted and used in different ways by different providers. This combination of simplicity and versatility is both a strength and a weakness. For instance, among the 15 basic fields

there is no placeholder for the name of the institution which awarded a thesis. Some providers use the dc:publisher element for this, others do not. Then, when it is present, an awarding institution may appear as [school, institution], [institution, school]; or just [institution] or [school]; with or without leading definite articles. Such variance between sources immediately poses difficulties for any third-party service, for which consistency is desirable. Finally, it is also common for meaning to be lost in the conversion from local data to a simple DC rendition of that information. For instance, an e-thesis will typically have several dates associated with it, such as date of approval, date of upload to the repository, perhaps date of digitisation; but wherever these are passed for export in OAI_DC format, they simply appear as a list of repeated dc:date entries. A DC harvester does not 'know' what any of these dates mean, nor which date in the list, if any, should be associated with the award of the thesis.

Dublin Core, therefore, brings accessibility, but at the expense of consistency. Unfortunately, a service which aspires to add value to harvested data, such as DART-Europe's Portal, is compromised without standardisation and consistency in its source data. This is, of course, a recognised problem, and — to generalise — there are two ways of addressing it. One is to set precise guidelines for the use of DC in any given context, and to enforce them rigorously. The other is to accept that post-harvesting intervention is likely to be required on the part of the aggregator. This is the route which DART-Europe has taken. The experience of building the DART Portal has suggested that many repository managers do not have the time or the resources to 'tune' their repositories to meet detailed export guidelines for metadata. DART-Europe, therefore, encourages participation by placing only minimal metadata obligations on contributors, concerning only the few elements of DC that are relevant to the Portal service and ensuring that at least a core author/title/date search and retrieval service is supported; and, as part of the aggregation process, applies various mappings and post-processing routines to incoming data to make it as useful as possible to the community. Figure 2 shows an overview of the technical architecture of the DART-Europe Portal.

This is very much a pragmatic outlook, and it is arguably a labour-intensive way of managing complexity. The underlying philosophy to date has been that as long as each data provider is consistent in its idiosyncracies, DART-Europe need only work out how to address them once, and those inter-

Fig. 2: DART-Europe E-theses Portal: technical overview.



ventions are then repeatable against all future incoming data. Is this a sustainable solution, scalable to the whole of Europe? Time will tell. Working with consortia certainly helps to mitigate the difficulties, but nonetheless it seems likely that, in the longer term, DART-Europe will find it necessary to implement a European e-theses metadata application profile to underpin the Portal, perhaps developing plugins for the most popular repository

software platforms to support its adoption. Certainly DART-Europe is unlikely to be able to offer services which are any richer than those already being piloted without a supply of more detailed and more consistent metadata. However, by setting realistic barriers for participation in the Portal, it is felt that DART-Europe has been able to make significantly more content available to researchers than it would if it had enforced more stringent technical requirements from the outset. The decision to pursue the relatively quick acquisition of a critical mass of content, allowing the demonstration of a basic portal and permitting some exploration of the limitations of 'plain vanilla' harvesting, has been an important strategic building-block in helping DART-Europe to look forward to the future development of more sophisticated services, while establishing a useful academic resource along the way.

The Portal: Future Plans

In the short term, there are issues still to resolve with some of the features which depend on high-quality metadata, such as the facility to limit searches by language. These features will either be stabilised or withdrawn. Some user-orientated functions are in development, such as the ability to save and manipulate marked lists of records, and a 'latest additions' feature. Usage statistics are also being investigated. It is hoped that the Portal will move formally to a production service before the end of 2008. Thereafter, it will perhaps be timely to begin to appraise possible cost-recovery business models to provide ongoing support for the Portal as a service. Running costs depend to a large extent on the functionality that has been implemented, of course, and user studies to find out what the community might want from an e-thesis Portal may also be appropriate in determing the direction of future development work. Meanwhile, the new collaboration with DRIVER referred to above may change the outlook completely: if successful, it would offer the prospect of the migration of the Portal service from the current harvester-interface coupling to the DRIVER infrastructure, something which the DART-Europe Board would be interested to consider provided that no content or functionality were sacrificed in the transition.

DART-Europe: Other Activities

DART-Europe pursues a number of other activities besides its E-theses Portal. Metadata and digital preservation have already been mentioned in this paper. DART also has non-technical interests: early in 2008, GUIDE (Guiding Universities in Doctoral Theses) merged into DART-Europe as DART-GUIDE. DART-GUIDE will have a particular focus on the support of advocacy and training as they relate to e-theses. Earlier work carried out by GUIDE has included a survey of e-theses developments across Europe, by country, and a website which collates toolkits, guidelines and best practice, covering advocacy, metadata, legal issues, and lessons learned from major European e-theses projects. That work is available in the GUIDE section of the DART-Europe website. It is a potentially helpful resource for organisations and researchers who are interested in aspects of electronic theses.

Conclusion

As may be expected of a programme which depends on contributions of time and resources from its partners rather than on any external funding streams, DART-Europe's progress has tended to be incremental, rather than explosive. However, as DART-Europe enters its fourth year, the partnership is able to point to a solid body of achievement.

Interest in joining the partnership has continued to grow, and all the partners benefit from the opportunity to participate in DART-Europe as a networking forum. The DART-Europe E-theses Portal is on the point of being stabilised and marketed to the academic community, and new collaborations on digital preservation with the Koninklijke Bibliotheek/National Library of the Netherlands and on harvesting technologies and services with DRIVER, and the revival of GUIDE's work under the auspices of DART-GUIDE, should soon begin to deliver some interesting outcomes. There is much for the DART-Europe partners to look forward to in the coming months, and it is hoped that DART-Europe's work will continue to support Europe's research libraries, library consortia and researchers, and continue to facilitate improved worldwide access to European research theses.

Websites Referred to in the Text

DART-Europe website, http://www.dart-europe.eu/About

LIBER, Ligue des Bibliothèques Européennes de Recherche, http://www.libereurope.eu

NDLTD, Networked Digital Library of Theses and Dissertations, http://www.ndltd.org

DART-Europe E-theses Portal, http://www.dart-europe.eu

GUIDE, Guiding Universities In Doctoral E-theses, http://www.dart-europe.eu/guide

EthOS, Electronic Theses Online System, http://www.ethos.ac.uk

DART-Europe Board, http://www.dart-europe.eu/About/contacts

DART-Europe Partners, http://www.dart-europe.eu/About/partners

Koninklijke Bibliotheek, http://www.kb.nl

DRIVER, Digital Repository Infrastructure Vision for European Research, http://www.driver-repository.eu/

PKP, Public Knowledge Project, http://pkp.sfu.ca/

Notes

[1] DART-Europe Partnership Agreement available at http://www.dart-europe.eu/About/documents/DART-Europe_Partnership_Agreement.pdf

[2] Memorandum of Understanding between LIBER and the Koninklijke Bibliotheek available at http://www.libereurope.eu/files/MOA%20LIBER%20KB.pdf

[3] DRIVER Guidelines for Content Providers available at http://www.driver-support.eu/documents/DRIVER_guidelines_1%200.pdf

[4] OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) full protocol available at http://www.openarchives.org/OAI/openarchivesprotocol.html