ETD 2008

Time to Harvest: Electronic Doctoral Theses in Italy

Stefania Arabito,[a] Daniela Cermesoni,[b] Paola Galimberti,[c] and Marialaura Vignocchi,[d]
(on behalf of the CRUI Working Group on Open Access, Italy)
[a]Sezione Ricerca e Dottorati, Università degli Studi di Trieste, Trieste, Italy; [b]Sistema Bibliotecario di Ateneo, Università degli Studi dell'Insubria, Varese, Italy; [c]Biblioteca di Scienze dell'Antichità e Filologia Moderna, Università degli Studi di Milano, Milano, Italy; [d]CIB, Alma Mater Studiorum, Università di Bologna, Bologna, Italy.
[a]arabito@units.it; [b]daniela.cermesoni@uninsubria.it;
[c]paola.galimberti@unimi.it; [d]vignocchi@cib.unibo.it

The Libraries Committee of the Conference of the Rectors of Italian Universities, hence CRUI, has recently approved the Guidelines for archiving doctoral theses in Institutional Repositories. The Guidelines are the first step of an initiative aimed at putting the principles of the Berlin Declaration into effect in Italian Universities. The CRUI Working Group on Open Access has conceived the Guidelines as a toolkit for Italian Universities, i.e. practical and legal advice for managing and disseminating theses via Open Access IRs.
This paper will detail the text of the guidelines reporting the main issues addressed by the Working Group. Legal implications have been rated as a top priority, and an embargo period has been required to protect patents and works in publication. Metadata have been defined in accordance with both European recommendations (Knowledge Exchange) and Italian National Libraries requirements, in order to implement national and international service interoperability. Delivery formats for long term preservation have likewise been judged as a matter of great importance.
This paper will illustrate how CRUI recommendations are affecting Italian University policies by presenting the results of a survey conducted early this year. The legal deposit of electronic doctoral theses via OAI-PMH, a parallel project started in 2007, has recently obtained full support by the Ministry of Cultural Heritage and has already reached the test phase. This paper will show the potential impact in terms of enhancing national discovery and provision services.

Introduction: the rationale for archiving doctoral theses in OA Institutional Repositories[1]

In Italy, the two National Libraries in Florence[2] and Rome[3] have to preserve all doctoral theses in paper and ensure public access on library premises by law[4]; the following restraints however cannot be overlooked:

- shipment from Universities and cataloguing procedures are extremely time consuming and delay the provision of physical access to doctoral theses;
- as doctoral theses are for library use only, they cannot be checked out nor can they be requested on interlibrary loan;
- no photocopying services are provided on premises.

As a rule, Universities have been storing doctoral theses only in paper. As for bibliographic records, they are often searchable in local University OPACs; in some cases they are only listed in local print catalogues. Italian doctoral theses have consequently had very little visibility so far; they are either hidden in the deep web or utterly missing from the web. On the contrary, doctoral theses are research outputs whose value should be recognized and enhanced through:

- immediate deposit in IRs, to the advantage of both post graduates and the institutions they belong to;
- Open Access[5] availability in compliance with the Berlin Declaration and the recent recommendations of the European Commission[6].

Doctoral theses have accordingly been included in the agenda of the Italian Working Group on Open Access[7], set up in April 2006 within the Conference of the Rectors of Italian Universities[8] Library Committee and chaired by Rector Vincenzo Milanesi. The OA WG aimed to implement the principles of the Berlin Declaration[9]. The starting assumption was to ensure visibility, dissemination and impact to doctoral theses by archiving them into OA IRs, as the first step towards establishing OA policies for all research outputs.

In fact, publishers are not involved in the validation of doctoral theses, as is the case with scholarly articles. Moreover young researchers become aware of the opportunities offered by OA publishing compared to traditional print publishing at the very outset of their academic careers. It is also a chance to instruct them not to sign their copyright away. Post graduates must learn how to retain the right to reuse and distribute their work by negotiating non-exclusive terms with commercial publishers.

The OA WG has accordingly developed specific guidelines and recommendations on doctoral theses targeted at Italian Universities. The guidelines were approved by the CRUI in November 2007.

The Guidelines

The ***Linee guida per il deposito delle tesi di dottorato negli archivi aperti*** *(Guidelines for archiving doctoral theses in IRs)*[10] are meant to promote best practices for capturing, storing, and disseminating electronic doctoral theses. A first survey was carried out in 2006 to outline common practices in Italian Universities; the outcome highlighted a diversity of situations; some Universities collected digital theses and made them OA available, with very few restrictions; some Universities had not even envisaged such possibility.[11]

The *Guidelines* take advantage of the models provided by Germany, The Netherlands, Great Britain, Denmark and Sweden. In these countries harvesting services have been established at a national level and have carried out the *European E-Theses* project, whose goal was to set up a portal giving access to doctoral theses at a European level.

The Guidelines aim at supplying a practical toolkit for all Italian Universities planning to deposit doctoral theses in IRs. Legal forms and clauses are included to make the necessary adjustments to the institutional rules regulating PhD courses consistent at a national level. A metadata scheme has been devised for the sake of interoperability. The main purpose is to simplify both administrative and technical procedures by offering practicable solutions.

Legal issues

The legal framework regulating doctoral theses in Italy is made up of laws pertaining to different domains, and serving opposite purposes. It is therefore difficult to steer clear of potential conflicts.

According to the Italian copyright law, the PhD student is the only author and as such holds all moral rights and economic exploitation rights. Hence, s/he has the right to prevent his/her thesis from being publicly available.

The laws regulating PhD courses[12] conversely hold BNCF responsible for making all PhD theses publicly available.

There is no specific law catering for electronic material and copyright issues, though. In fact most Universities do not provide full access to e-theses lest copyright infringements may ensue. These Universities have not set up an IR yet.

However, the assumption of the OA WG is that accessibility should be nation-wide and by no means institution-wide.

Institutional Repositories

According to the CRUI recommendations, Italian Universities have the right and the duty to mandate self-archiving of doctoral theses in their IRs, on the assumption that doctoral theses become public as soon as they are defended. Nonetheless the institutional policies regulating PhD courses have to be suitably modified. When enrolling for a PhD, post graduate students will be notified that their theses will be made OA available in the IR after defence as part of the requirements for being granted a PhD degree. They will have to comply with such provision but for a few exceptions where an embargo period will be allowed (pending publication or patenting, third party agreements).

For ongoing courses immediate deposit can be mandated. PhD students will be asked if they want their theses to be OA or not. If not they can ask for an embargo period ranging from six months to three years.

All universities should adopt similar strategies and go for self-archiving instead of mediated deposit by librarians, in order to make authors responsible for the integrity of their work. All doctoral theses will be submitted to, or harvested by, BNCF and BNCR, which are in charge of legal deposit, long-term preservation and national discovery and accessibility.

Integration and interoperability require the adoption of standard protocols and metadata.

Copyright clearance and embargo

As mentioned above, the free availability of doctoral theses on the web can be jeopardized by thorny copyright issues, which arise in the following cases:

- use of third party owned materials (if they are copyrighted, or if the research is funded by external agencies, a written permission is required unless otherwise agreed) [13]
- third parties involved (possible infringement of privacy)
- patentable discoveries (even though PhD students should apply for patents before defending their theses)
- ongoing publication of data (according to the publisher policy)

It is therefore sensible, nay indispensable, to allow for an embargo period, in compliance with the immediate deposit/optional access model[14]. PhD students are required to self archive both the metadata and full text theses; the metadata will be immediately searchable and retrievable, while the e-theses will be embargoed for a period ranging from six to twelve months.

Sharing the metadata

The recommendations of the OA WG started from a comparative analysis between the metadata sets commonly used in Italian IRs and current European practices[15]. The outcome of this work is a metadata set attached to the Guidelines. The purpose is to have interoperable repositories not only in Italy but also in Europe and to share standardized procedures.

True enough, all the main harvesting servers can convert the metadata set of an IR to a standard format. It is crucial however to build upon a common metadata set, in order to avoid subsequent interventions to normalize metadata both within and outside Italy.

The levels of interoperability range from harvesting all full text doctoral theses in IRs to building value added services for doctoral theses on top of repositories.

Presently, the only viable objective is to attain the first interoperability level, by sharing a common protocol for data exchange (OAI-PMH), information exchange and data structure, with an accurate definition of the meanings of all fields.

Simple Dublin Core has been considered inaccurate for an advanced search, which is conversely ensured by a DC qualified metadata set. Some fields are mandatory, others are recommended or optional. Priority fields are listed below:

dc.title: title of the work;
dc.creator: author of the work (surname, name);
dc.description: abstract (better if in English);
dc.language: language (format ISO639-1);
dc.identifier: URL of the thesis full-text or of a halfway page;
dc.type: Doctoral Thesis (only in English);
dc.contributor: tutor/supervisor (surname, name);
dc.date: date of publication (ISO 8601), i.e. date of defence; this is the only date in metadata; other dates in other fields may be misleading;
dc.publisher: name of the University
dc.format: dimension in bytes/MIME type.

Other fields:

dc.subject: classification of subject fields according to the Ministry of Education and Research;
dc.rights: embargo or immediate availability

The Italian repositories complying with this metadata set will be compatible with European standards. Further steps may entail a collective definition of the broader context of the information exchange, such as a standardized assessment of PhD degrees in every country.

File formats

Choosing the “right” file formats is still a controversial issue, especially in the domain of e-theses. Theses are at the same time bibliographic items and administrative documents and therefore have to comply with multiple requisites.

As administrative documents, their authenticity, integrity and fixity should be secure and enduring for the sake of long-term digital preservation. They should also be assigned suitable archival metadata.

As bibliographic items, they must obey to the criteria of web accessibility. Actually, they are produced by young researchers who have no domain expertise in the field of digital curation and who have a limited informatics toolset, typically restricted to word processing and basic filing. PhD students are not aware of format obsolescence, but they are interested in protecting their work from unintentional/malicious altering after web publishing. However, files must not be encrypted, in order to permit refreshing.

The best approach would probably be to educate both undergraduates and post graduates on long-term preservation issues and to provide them with the correct tools to produce xml files which embed bibliographical and archival metadata. Archival copies and web publications would then follow quite naturally. Needless to say, liaising with faculty is a key factor; academics should also be granted user-friendly and effective authoring tools, ideally invisible to them.

However, a nation-wide strategy is needed; all the parties involved, namely archivists and the National Libraries, should be called to share their expertise in the field of long-term preservation. Indeed, Italian archivists are presently developing a specific set of metadata, while the National Libraries are the institutions appointed by law to preserve and curate the Italian bibliographic output.

It is impossible at this stage to settle the tricky and thorny issue of file formats. Still Italian Universities needed to start collecting and storing their e-theses. PDF/A has eventually been chosen – in accordance with the requirements of the National Libraries. It has not been a trouble-free and straightforward choice, though. PDF is often considered anathema – whoever attended the latest Open Repositories meeting in Southampton will know about it – on the grounds that it does not allow text mining. On the other hand, PDF has become a de facto standard all over Europe and beyond (apart from Germany). The recent Digital Preservation Coalition recommendations[16] corroborate this position; yet, the matter is not to be considered settled at all, and a constant evaluation of file formats is required.

IRs and National Libraries

In July 2007, the Ministry of Education and Research sent a circular note to all Universities stating that the legal deposit of doctoral theses could henceforth rely on digital technology and that paper copies were no longer needed[17]. The OA WG immediately started a project with BNCF and BNCR to test the feasibility of the new system. The project intends to analyze the workflow of the legal deposit of e-theses and to implement the necessary technological infrastructure to automate the procedure.

As mentioned above, the discovery and access services to doctoral theses provided by BNCF and BNCR have hitherto proved to be both untimely and uneconomical. The national output amounts to roughly 9,000 doctoral theses per year[18]. Before the Ministry provision, paper copies of all doctoral theses were sent by the Universities both to BNCF and to BNCR. They were then catalogued and indexed, with a waste of time and resources. It has been estimated that bibliographic records are available on average four years after theses have been defended. Delayed public access is justified by such a large output as opposed to the limited resources of BNCF and BNCR. Moreover, Universities are generally late in shipping the paper copies of doctoral theses.

Many advantages are expected from the project of the electronic legal deposit of doctoral theses. First, it will improve the related national bibliographic services, offering a quicker update of the on-line catalogue and eventually direct access to the full-text. It will also simplify administrative and bibliographic procedures both within the National Libraries and

within University administrations, with a consequent cutback on costs. The project has also stirred up a pragmatic national debate on metadata and file formats which will hopefully attract all the stakeholders, with a bit of luck the archivist community.

The project has obtained full support from the Ministry of the Cultural Heritage and may spur urge prompt the long-awaited reform of the roles the two National Libraries, in terms of better definition and allocation of resources and functions. In this case, BNCF has provided the technological expertise and infrastructure, while BNCR has contributed to the analysis and will be one the three back up sites required for trusted digital repositories.

The whole workflow has been studied by the National Libraries together with representatives of the OA WG. At this early stage, two possible technical procedures have been put forward, in order to enable all Universities to join.

BNCF will harvest both the metadata and full-texts via OAI-PMH of the doctoral theses of the Universities which have already implemented their IRs.

An alternative upload via web form will be put in place for the Universities which have not set up an OAI-PMH compliant repository yet. Once BNCF has either harvested or received the metadata and full-text theses, it will send SHA1 hashes back to Universities to certify the deposit. Metadata will be then validated by the librarians in charge of the Italian National Bibliography and finally imported into the BNCF OPAC.

The planned workflow has only been partially accomplished. Harvesting via OAI-PMH has already been tested successfully with the IR of the University of Bologna, but SHA1 hashes have not been sent back and metadata have not been validated. Actually, even the metadata issue has not been settled yet. At this stage BNCF required simple and not qualified Dublin Core. The National Libraries and Universities will have to jointly develop functional metadata for legal deposit procedures, including the automated management of digital rights.

E-theses and IRs in Italy: work in progress

Last January the OA WG carried out a survey amongst Italian Universities. Here are the main outcomes of the questionnaire:

- 25 Universities were collecting or about to collect electronic doctoral theses in IRs, mostly to make them OA available;
- the local output ranges from 50 to 500-800 doctoral theses per year, according to the size of the institutions, with an average of 200-300 theses per year;
- only a few Academic Senates have officially mandated the deposit of doctoral theses in IRs; 50% put a mandate on the deposit only, 50% put a mandate on OA availability;
- librarians mainly advocated depositing in the IR;
- workflow procedures were taken in charge mainly by librarians, with the cooperation of administrative staff;
- DSpace and Eprints are the most popular software tools. IRs are in most cases integrated with other databases, particularly administrative databases, authentication systems, research archives, OPACs, cross-search utilities;
- self-archiving is a common ingestion procedure; in most cases however librarians deposit e-theses;
- embargo is usually allowed for periods ranging from six months to three years. Twelve months is the standard embargo period for almost 50% of the Universities. The length of the embargo is very rarely left to the choice of PhD students. In all cases they are required to sign a declaration stating the reasons for the embargo. In all cases metadata are immediately searchable and retrievable;

- services added to basic e-theses management (print on demand, legal deposit, statistical services, legal advice for users, permanent preservation) are supplied by 50% of the institutions; 50% is planning to implement them;
- libraries grant financial support in 50% of the Universities; research and/or other institutional units contribute to maintain 50% of IRs;
- the workflow relies on the interaction between librarians, computer specialists, administrative staff, and tutors. Seldom are librarians the only actors involved;
- half IRs are dependent on outsourcing contracts, half are managed with internal resources;
- all Universities would appreciate nation-wide interoperability, namely on syntactic (federated search on multiple archives) and semantic (multilingual, discipline and subject search) bases. Semantic interoperability is considered to be very hard to achieve, though. A dedicated Italian harvester would be top-priority.

According to the surveyed institutions, the following issues should be tackled at a European level to enhance the value of electronic doctoral theses:

- widespread archiving programmes;
- common standards;
- a European portal/network for doctoral theses;
- ongoing advocacy;
- joint international projects;
- joint participation to international conferences;
- equivalent systems of higher education in Italy and in Europe;
- economic, legal and technical support;
- unlocking the potential of PhD research;
- networking with publishers.

E-theses and IRs in Italy: two case studies

1. Bologna

The case of the e-theses project at the University of Bologna epitomizes that times are ripe for changing the way doctoral theses are stored and disseminated - as the recent OA WG survey bears out. Launched in 2006, the project aims at setting up an OAI-PMH compliant IR in order to collect, organize and provide access to the doctoral theses produced at the University of Bologna, no less than 650/800 per year[19]. No bibliographic services for doctoral theses had ever been actually provided before. Now the repository stores, indexes and provides access to all the theses defended in 2007 and 2008, and is harvested by the BNCF for legal deposit.

The case of Bologna University, however, also reveals the criticalities and weaknesses that need to be addressed and settled to radically reform the global context by really exploiting the opportunities offered by the digital environment to their fullest potential. On a technical level the project relies on a well equipped digital library infrastructure and on the expertise of well trained staff, which made the implementation of the repository and the redesign of the workflow timely and effective. The new procedure for legal deposit has been approved by the administration. Yet, the project has not evolved into a fully legitimate institutional procedure yet.

The University assessment body has stated that publishing doctoral theses in OA is a valid criterion for evaluating doctoral schools. Nonetheless, the various scientific communities have reacted to this indication differently. The needs and expectations that have emerged vary according to the disciplines. Impact factor heavily influences research outputs evaluation in

certain fields. Young researchers accordingly reprocess their doctoral theses as articles and submit them to journals with an IF. Researchers in humanities and legal studies also show resistance to an OA distribution of their theses. The reason is that they normally transform their doctoral theses in their first monographic publication. This is why Bologna University has not put a true mandate on OA publication for doctoral theses yet. The repository however contains only a small number of documents that are actually not accessible at all and not merely embargoed.

A statistical analysis of the levels of access chosen by doctoral students shows a good acceptance of OA[20]. The most common reasons for restricting the availability of their theses proves once again how the commercial publishing system is deeply rooted in the academic domain. OA is not actually rejected, it is rather considered to be a sort of second best choice.

What we have hitherto observed pushes us to focus on the different needs of the various research communities. We will have to find better and new ways to make the repository more appealing in order to transform it into a real service supporting research.

2. Trieste

Trieste is a medium-sized University with 27,000 students (both undergraduates and postgraduates), twelve faculties and over forty research departments. The annual output of doctoral theses is roughly 200 per year. The Library System has always been in charge of storing, cataloguing and providing access to theses in paper, including doctoral theses, acting as the Registrar's Department archive. Data on doctoral theses were entered both by the Library System and the Registrar's Department in different and independent databases. Paper theses occupy miles and miles of shelves in a jam packed warehouse distant from the campus. They can be searched in the OPAC – after time-consuming and slow cataloguing - and then accessed with considerable delay.

OpenstarTs [21]started as an institutional project managed by the Library System in 2006. It aimed at granting theses higher visibility and at maximizing their impact by depositing them in an OAI-PMH compliant repository. Sustainability was a major constraint, as the Library System could not afford to allocate human resources but for the project manager (on a part-time basis). The project could cover only start-up expenses, namely the technical support of an external consultant; it was therefore crucial to take a "lean" approach through full cooperation with the PhD department and interoperability with existing databases.

To avoid unnecessary duplications and to simplify both the ingestion and the validation process, the repository was integrated with LDAP for authentication and with the Registrar's Department data warehouse for the relevant metadata. In 2007 the system was tested by PhD students who volunteered to self archive their theses. They appreciated the user friendliness of the procedure; apart from actually uploading their PDF, they were requested to enter only abstracts and keywords, as all the other metadata had already been entered (and validated) via the Registrar's Department data warehouse and subsequently mapped in the repository.

After customizing the DSpace workflow and adapting PhD regulations, self archiving doctoral theses in the repository was made compulsory as part of the requirements for defence and for the award of a PhD degree in 2008. Post graduate students were allowed to opt for a one-year embargo and asked to specify the reason for their request. Their feelings about OA publishing mirror what already observed at the University of Bologna.

The 2008 output amounts to 181 doctoral theses, which will soon be harvested by the National Library of Florence.

A few technical issues emerged, namely the upload of heavy files; the main concern however is to make the repository really relevant to the institution in order to find the resources to build more services on top of it and to tailor them effectively.

Conclusion

The aim of the OA WG was to publish the Guidelines, as a reference tool for the Universities planning to deposit doctoral theses in IRs and to make them OA available. Sharing the same metadata set was strongly recommended for the sake of interoperability with European portals.

The goal was the dissemination of doctoral theses, given their importance as research outputs and their lack of visibility on the web and consequent lost impact. The certification and preservation of doctoral theses are tasks which other bodies are in charge of and hold responsibility for.

The Guidelines have been greatly appreciated by all the surveyed institutions, as proved by the outcome of the survey. The Universities that had already set up an IR considered them important inasmuch as they justified and backed up the choice of having the repository. In most cases the top management was not totally aware of the potential of the repository, which was not considered to be a fully legitimate institutional tool. Wherever the deposit of doctoral theses was at a start-up stage, the Guidelines provided support to the project and justified the request for an institutional OA policy. In all cases, the Guidelines represent a milestone for putting into effect a national common strategy on OA.

Needless to say, the mission of the OA WG has not been accomplished yet. The first step has been taken, but pervasive advocacy is needed to expand existing IRs and to neutralize resistance to change. Opposition to the deposit of electronic doctoral theses in IRs and to their OA availability is still strong and widespread.

It will also be important to keep constantly and closely in touch with other European and international working groups to monitor their progress and keep up with their activities.

Value added services for this kind of materials will have to be established at European level too: syntactic and semantic interoperability, the use of common standards, a dedicated Italian harvester are other items in the agenda.

It will also be vital to instruct PhD students on their rights as authors and on the correct use of third party materials.

Last but not least, an exchange of ideas with archivists on the preservation and certification of doctoral theses is top priority.

References

[1] Hence, IRs.

[2] Hence, BNCF.

[3] Hence, BNCR.

[4] DPR 07/11/80, n. 382, art 73, D.M. n. 224 04/30/99.

[5] Hence, OA.

[6] http://cordis.europa.eu/search/index.cfm?fuseaction=news.document&N_RCN=29243

[7] http://www.crui.it/HomePage.aspx?ref=894 hence, OA WG.

[8] hence CRUI.

[9] http://oa.mpg.de/openaccess-berlin/berlin_declaration.pdf; http://oa.mpg.de/openaccess-berlin/signatories.html

[10] http://www.crui.it/HomePage.aspx?ref=1149

[11] http://www.ukoln.ac.uk/repositories/european-e-theses/index/Italy

[12] DPR 07/11/80, n. 382, art 73, D.M. n. 224 04/30/99.

[13] The OA WG has specifically dealt with the management of all kinds of third parties materials (images, photos etc.).

[14] Harnad, Stevan http://eprints.ecs.soton.ac.uk/14431/4/pehm-harnad.pdf

[15] http://www.knowledge-exchange.info/

[16] Betsy A. Fanning, *Preserving the Data Explosion: Using PDF*. DPC, 2008. www.dpconline.org/docs/reports/dpctw08-02.pdf

[17] Circolare n. 1746 del 20 luglio 2007.

[18] The number of doctoral theses in Italy increases every year: from 4,000 defences in 2000, to 9,800 defences in 2005.

[19] http://amsdottorato.cib.unibo.it/

[20] The theses already published are 565 and 7% only are subjected to availability restrictions.

[21] http://www.openstarts.units.it