# Dissertations as Data

Joachim Schöpfel, Eric Kergosien, Stéphane Chaudiron, Bernard Jacquemin

# Dissertation as Data

Joachim SCHöPFEL, Stéphane CHAUDIRON
Éric KERGOSIEN, Bernard JACQUEMIN

## Introduction

Being originally handwritten by students and exclusively intended for teachers and professors, academic theses and dissertations have dramatically evolved in form, content and usages over the past century. By the beginning of the 20th century, students began to have their writings typed using mechanic typewriter, conferring to the works a more professional status. Those documents became stored and catalogued in university libraries, and read or even used by academic communities for research purpose.

A new step has been taken in the 90s when it became common for most of the students to type their works by themselves and directly on a personal computer. Even if those writings were (and very often still are) submitted in a paper format, they are initially produced in an electronic format, and can be stored and archived in that way. They represent the starting point of electronic theses and dissertations (ETDs). And despite the academic habits to deal with material writings, the mere existence of those numeric files, increasingly made available by the authors regularly driven by their academic institutions, has major consequences on both document processing and the way they may be taped by researchers. In particular, ETDs have obviously developed functional features specific to the three-fold nature of electronic documents studied by [Pédauque, 2006]. The digital nature of ETDs, established by the distinction between materiality (form), content (sign) and social context (medium), offers inchoately new ways to take advantage of those them.

## Dissertation as a document

According to [Pédauque, 2006], a document has a triple reality, inseparable in paperwork, but that can be discriminated in electronic documents such as ETDs. In order to evaluate consequences of the digital transition, we first observe the pre-digital theses and dissertations in accordance to the same criteria as we will later for ETDs.

Regarding the physicality of the dissertations (form), the material aspect of the works has to be taken into account. Traditional dissertations have format characteristics (e.g number of pages, paper size, font style and size), several media to be hold (e.g paper, microfilm, microfiche), internal structure rather formal but that may differ from an institution or a scientific field to another (e.g IRMED in experimental sciences, bibliography), they may include figures or charts, and sometimes annexes and other research materials such as corpora or recordings.

The meaning of the textual contents on the physical support (sign) basically relies upon lexical semantics. Consequently, correct access to knowledge or even to information into the dissertations can be provided by considering two types of contexts: the scientific context (domain, bibliographic references) of the dissertation determines a semantic and terminological field where word meaning can be selected; and the lexical and syntactic context in which each term appears also applies semantic constraints.

Finally, the dissertation is written and published in a purpose specific for a human society (medium): the author has to communicate his scientific approach and his achievements as a clue of abilities acquired along a doctoral program. The communication occurs in a social environment where it has to prove legitimacy, scientificity and authorship. For dissertations, the legitimacy is strongly attached to the institutional rules and processes of practicing science, writing and depositing a document, defending a scientific work before a jury, etc. The scientificity has to be measured by peers essentially in terms of publications and dissemination, or according to the way the contents are connected to previous research. As a creative work, a dissertation carries authorship for both text production and scientific development likely to have impact on technological and financial interests. Thus, dissertations have powerful links to the civil society.

## Dissertations and data

All along a research work performed in view of writing and defending a doctoral these, research material and data are collected, processed and preserved. Besides their own data, researchers have always based their work on previous results and often on data collected by others in other circumstances. For decades, "outside" data would have been used more extensively for new research, had they been made easily accessible and operable. The electronic property that characterizes today's dissertations potentially offers

technical ways to capture more directly research data. The current challenge is now to switch from the virtual to reality. Consequently, potentialities have to be identified. Several inquiries have been conducted these last years that make a significant overview of the situation, and two cases arise: research data may be included into the textual document, or the dissertation may provide links to research data, that are stored somewhere else.

*Dissertations as a data vehicle*

Including research material directly into the document is the most obvious way for dissertations to provide access to data. Several inquiries have been conducted among 780 print and electronic dissertations in sciences, social sciences and humanities from the Universities of Lille (France) and Ljubljana (Slovenia), submitted and defended between 1987 and 2015 [Prost et al., 2015,Schöpfel et al., 2016]. Those inquiries have been able to distinguish between different aspects of data, considering the primary (raw data from field practice, sources) or secondary (processed information, enriched results) nature, or taking into account if they appear into the dissertation text or attached to the text as an appendix. Results are also observed depending on the academic disciplines of the dissertations.

When dissertations contain research data (526), the chance for an external user to reuse them for other purpose depends strongly on those parameters. Primary data consist mostly of interviews or surveys, experimental observations gathering or text samples (including scanned archives). As source of the research approach, those pieces are broad, and can be reused in many contexts, assuming that the information format allows it, which is not obvious for each format. Secondary data are obtained through the research operations, as a result of the source processing. Often more structured or more elaborated (tables and statistics, tagged text corpora, charts and drawings), the results are likely to be reused. Nonetheless the reusability is linked to documentation and to other factors: figures and tables into the dissertation text are given a caption, but remain difficult to be harvested properly by a computer tool; in appendices, results or even sources are less completely documented, not to mention the lack of metadata or simply of an unique identifier, and the large variety of formats and technical media makes them incompatible with content mining algorithms.

According to the inquiries, habits of communicating research data directly in dissertations prove to depend strongly on the scientific domain. For example, many dissertations in historic science give access to large amount of data, in a wide variety of formats, including structured and reusable ones. If archaeology and history of arts dissertations providing data are fewer, the quantity of data supplied is significant, but more concentrated with iconography. In contrast, philosophy tends to communicate research data, but in small volumes, whereas political science deals with few data in few documents.

*Dissertations as a gateway to data*

Another way for dissertations to give access to materials used for a research project is to provide link to primary or secondary datasets located elsewhere than into the document. Dissertation text and research data can be described and stored apart, in distinct repositories. Depending on situations, data can be located on either institutional or personal repositories, and if applicable, corresponding ETDs are uploaded to other specific facilities. Data formats can vary depending on the publication aim (data preservation, dissemination or even initial gathering to be shared between actors): institutional repositories have requirements for a rather structured information, and are often more demanding on this topic than social networks. The breakup of the dissertations into text and data creates a need: links have to be established between the documents and corresponding data; they may be provided by means of a (preferably permanent) URL or a unique data identifier determined in or outside the scientific domain (DOI).

Several projects have studied actual cases of separation of text and data stored on different platforms in order to introduce workflow solutions or process recommendations. The Dataverse ETD program conducted at Emory University [Doty et al., 2015] involves librarians to identify precisely dissertations submitted with supplemental files containing research data. They download and document research data in Dataverse repository, and establish an explicit link between the record in ETD system (for a given dissertation) and the record in the data archiving service (for the corresponding research data). ETDplus project carried out at Educopia Institute has leaded to guidelines that takes a similar position [Schultz et al., 2014]: research data should be stored in a repository separate from the dissertation text, and a permanent link between text and data should be provided into metadata by using persistent identifier such as handle,

DOI, ARK or PURL. Furthermore, data files have to be processed to match archival format requirements. Bielefeld University has set up PUB, a functional workflow originally designed to provide a bridge between research data and working papers, and also able to cope with digital dissertations that may be accompanied with research material in additional files [Vompras and Schirrwagen, 2015]. The working paper/dissertation and the data files are stored separately in respectively a disciplinary repository and an institutional repository. A DOI is assigned to each object according to the DataCite Metadata process, that ensure the linking between text and data by both DOI citations in the text and link explicitation in the DataCite Metadata Schema.

*Dissertations as data*

The approach we defend here differs from the previous ones by considering dissertations themselves as data. No distinction is made anymore between dissertation text and attached data, as the whole content is regarded the same way by content mining: structured data can be seized by the process, as contents can be collected, measured and analysed to provide useful information. Even metadata associated to dissertations and to research materials can supply a data mining system.

The use of data mining techniques on dissertations entails that source dissertations are in digital format. The digital condition of documents is constitutive of considering ETDs as data, and to make it possible data to be processed by digital tools.

Considering the document theory stated by [Pédauque, 2006], a document is comprised of three dimensions: form, sign and medium. In the digital world, those dimensions have evolved: the paper document's form relies on an inscription performed on a medium, but digital documents dissociates inscription (stable) and supports (that can vary from writing to reading or storage). Thus electronic dissertation's form consists in a structure in which are organized data. Being expressed according to a specific language (XML, TEI), to an internal (DTD) and external (RDF) structure for a digital support (web), data fit information extraction and filtering systems.

The paper sign has particular interest in the way that a content constructs the meaning. ETDs sign uses structured data to build knowledge, by using both lexical and semantic hints, and external information resources such as ontologies, DBpedia, geolocation or specialized vocabularies. Automated systems can identify (and virtually reuse) arguments and finality of ETDs.

In physical documents, the medium dimension includes the notion of social communication and an accurate context specified by rules or even laws that legitimate the way the communication is performed. Transposition in ETDs implicates strong rules that prescript how dissertation or additional material has to be structured, which format is demanded, and what procedures are involved. Moreover, legal status is attached to ETDs, indicating precisely access permissions or restrictions to the document, licensing information and possible intellectual property specifications when applicable.

## Using dissertations as data

While researchers and institutions are well aware of the interest and the need to organize storage and availability of ETDs and research data, very few projects have been undertaken to use those very files as research data.

*SPECTRa-T*

Yet the SPECTRa-T project [Morgan et al., 2008] has designed a set of tools to perform text mining on chemistry ETDs and extract research data in order to build RDF triplestores and enable semantic queries. This proof-of-concept approach addresses issues attached to the three dimensions of the digital document. Although data mining tools are more efficient when applied on structured text, such as marked-up documents formatted in XML, MS Office Open XML (.docx) or OpenDocument XML (.odt), ETDs are often supplied in other, less suitable formats: PDF but also PostScript, LaTeX and MS Document Format (.doc). To struggle against this heterogeneity of form, in the absence of a marked-up document file, the process requires a modifiable Word version instead of PDF or a PostScript, in order to convert it to XML as a standard processing format (form).

The access to meaning is granted by text and data mining tools, specialized vocabularies or ontologies (sign). These tools and resources can only be efficient if they are designed according to the advice of subject experts in the same discipline. Even more, the tools and resources implicated in the workflow have to be dedicated for an individual discipline. Repositories involved in data storage and especially in

triplestores building offer services and processes that meet the needs of researchers of the field and that match the domain practices: those repositories are designed by working closely with researchers in different subjects.
As scientific communication supports that legitimate both research skills and intellectual property (medium), ETDs have to be both preserved for the archival role of assessing students' research, and valued by the institutions responsible for the production (e.g universities) an the conservation (e.g repositories), because they are unique resources containing potentially valuable data that must be made extractable and re-usable. In order to overtake a simple archival process and foresee new research purposes, the whole approach of producing, depositing and processing ETDs has to be designed to ensure the chance to reuse. Discipline-specific policies have to be published to manage each step of the procedure and to match ETDs processed contents with data extraction tools and with new needs of research.
SPECTRa-T has designed a workflow and guidelines to deal with heterogeneous ETDs formats and make it certain that the files are converted into standard XML, to take into account chemistry domain and sub-domains while lexico-semantic resources and tools extract data, store them into repositories and build semantic links in triplestores. In this approach, technical issues and modelling seem to can be addressed: data can be seized by content mining algorithms, and information extraction leads to knowledge databases. [Brook et al., 2014] note that the remaining gap lays in the choices to limit the access to ETDs and databases, either related to legal issues around copyright and database rights, or connected to publishers' restriction to physically access data. Those limitations have to be solved in a non-technical field.

*TERRE-ISTEX*

This ongoing project[1] proposes an approach to identify and extract information (particularly geographic data) from scientific texts in order to develop an information research tool. TERRE-ISTEX is part of a larger project, ISTEX[2] - one of the "Investment for the future" program of the French Ministry for Education and Research -, whose main objective is to provide free and easy online access to retrospective collections of sciences literature in all disciplines, including scientific journals, databases, text corpora, etc. Being a subpart of this project, TERRE-ISTEX ambition is to identify, in a large heterogeneous corpus, named entities, geographic and thematic information linked to climate change.
Our corpus is basically constituted of more than 8 million scientific articles from ISTEX data with linked metadata in MODS format, of 25,000 multilingual documents (articles and other grey literature texts) from CIRAD, and of 70,000 digital or digitalized dissertations and metadata from ANRT. Among those documents, only a small part is relevant for the climate change subject, but they have to be identified by the process. The first step of the general process is thus to automatically identify in this large, heterogeneous and multilingual corpus thematic, location and temporal data. After a semi-automatic validation performed by subject experts, data are indexed in the Elasticsearch[3] tool in order to plan search operations among identified information in plain text, and to enable geographic content/results analysis.
As a proof-of-concept, we first processed 1,200 articles from the EGC conferences[4], and we built an index from the relevant information automatically identified in those texts, including thematic, spatial, temporal and structural information. The NLP tool gathered data such as conference place, author city names (that can be linked to place or city geocodes), titles, abstracts, domains, author names, year of the conference... The automatic data-markup validation is performed in both French and English by thematic experts using the SentiAnnotator viewer [Farvardin et al., 2015], and the results can be stored in the Elasticsearch index. The analysis of the markups in Elasticsearch provide first interesting results: the thematic layout along the period under review indicates the evolution of relative thematic significance of the domain, assuming that more publications are accepted in an important topic; the thematic evolution can also be linked to locations of interest, considering the research laboratories and the conference places; relations of co-

1 https://terreistex.hypotheses.org/.
2 http://www.istex.fr/.
3 https://www.elastic.co/products/elasticsearch.
4 http://www.egc.asso.fr.

authorship and evolution of the co-authorship along the period has also been under review, linked or not to the identified topics[5].

From this first experimentation, we prospect now to scale up to the "big data", taking into account the heterogeneity (languages, formats, text typology) of the corpora, and to test other scientific domains like history, culture... We also need to integrate other analysis resources such as knowledge database, web contents, Open Archives... to narrow the results.

**Barriers**

For an easy operation of dissertations in a research perspective, several bolts - technical, psychological, legal, commercial... - have still to be unlocked.

Availability: The text mining tools can process only electronic text or data. One part of dissertations, especially the older, do not exist in a digital format. Moreover, digitization of printed dissertations is not yet a solution, producing often inoperable images or little more usable document format such as PDF.

Accessibility: An access to dissertation should be provided to allow their reusability. One part of - even digital - dissertations are not made online available. And offline dissertations are alas mostly difficult to identify, and thus to find. Even when a digital documents exist, few of them cannot be shared because of confidentiality; a little more are embargoed according to a publisher contract because they are to be published as a book; and the other have access restrictions settled by a repository platform. Sometimes, non-institutional sharing is performed on a personal website or on social networks, making the dissertation difficult to find or to download.

Legality: Legal questions remain in several countries about access to ETDs, and more about text and data mining in dissertations: just because ETDs are online does not necessarily mean information processes can be applied, because no law or license has made a legal framework to scientific TDM. Another pending question involves remuneration for the authorship (sometimes difficult to identify) and for the storage of ETDs, not often take into account: open access is not equal to free access, and providing someone with the right to read a dissertation is different to grant him to apply TDM on that very dissertation.

Feasibility: Notwithstanding a positive view on text and data mining tools that can process online ETDs, those approaches are not optimal: the format (especially PDF) does not fit with marked up data, needed by analysis tools; the inner structure of text and data is often incorrectly identified by automated algorithms; metadata associated to ETDs could be of help, if comprehensive and provided by a discipline or domain expert also qualified in documentation, which is very rare; and above all, TDM tools and resources included in the process or the workflow have to be scientifically adapted to the discipline under review, preferably to a totally generic platform.

**Concluding remarks**

Under the Netherlands UE presidency, major actors of research (researchers), of research communication (commercial and open publishers) and of research and technology users (industrial partners) met in Amsterdam on April 2016. They together published a *Call for Actions on Open Science*[6] that prompts political, commercial and educational institutions to take up data mining in scientific document issues, and to initiate rules and actions to foster success in opening access to scientific documents for TDM. The C*all for Actions* propose 12 general measures that initiated the rallying of more specific claims: in October 2015, the LERU (League of European Research Universities) already demanded "a mandatory exception that will enable users to text and data mine all content to which they have legal access: the right to read is the right to mine"[7]. In France, a *Law for a Digital Republic*[8] has been published on October 8th, 2016, that institute an exception to copyright for text and data mining when scientific documents are legally accessed. This

---

5 First results, graphs and drawings are available at http://ekergosien.net/DefiEGC/index.html.

6 https://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science.

7 http://www.leru.org/index.php/public/news/the-right-to-read-is-the-right-to-mine/.

8 https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte.

exception is valid only for scientific purpose, and not for commercial use. The European Union should legalize in 2017 a similar exception for research and educational purpose, enabling text and data mining to texts and contents to which legal access is granted. Furthermore, publisher are asked to adapt both licensing and technical format to make mining easier.

The conditions of success of these initiatives can be summed up in four words: availability, accessibility, legality, feasibility (A2LF), that have to be reached to unlock current bolts. ETDs and additional materials have to be made available by both retro-digitization and deposit of text and additional data on online repositories. Access has to be granted by free and technical open access to clearly identified institutional repositories. The legality of mining ETDs has to be ensured by a legal TDM exception; moreover, the issuing of prescription rules should systematize a third party agreement to clear rights in a mining context. Prescription rules could also ease the feasibility by proposing application standards and by promoting rich metadata and text structures.

## Bibliography


[Brook et al., 2014] Brook, M., Murray-Rust, P., and Oppenheim, C. (2014). The Social, Political and Legal Aspects of Text and Data Mining (TDM). *D-Lib Magazine*, 20(11/12).

[Doty et al., 2015] Doty, J., Kowalski, M. T., Nash, B. C., and O'Riordan, S. F. (2015). Making Student Research Data Discoverable: A Pilot Program Using Dataverse. *Journal of Librarianship and Scholarly Communication*, 3(2):eP1234 1-25.

[Farvardin et al., 2015] Farvardin, M. A., Kergosien, E., Roche, M., and Teisseire, M. (2015). [Demo] A webtool for analyzing land-use planning documents. In Villata, S., Pan, J. Z., and Dragoni, M., editors, *Proceedings of the ISWC 2015 Posters & Demonstrations Track*, pages 1-4, Bethlehem (PA).

[Morgan et al., 2008] Morgan, P., Downing, J., Murray-Rust, P., Stewart, D., Tonge, A. P., Townsend, J., Harvey, M. J., and Rzepa, H. (2008). Extracting and re-using research data from chemistry e-theses: the SPECTRa-T project. In *11th International Symposium on Electronic Theses and Dissertations*, Robert Gordon University, Aberdeen.

[Prost et al., 2015] Prost, H., Malleret, C., and Schöpfel, J. (2015). Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities. *Journal of Librarianship and Scholarly Communication*, 3(2):eP1230 1-19.

[Pédauque, 2006] Pédauque, R. T. (2006). *Le document à la lumière du numérique*. C&F Éditions, Caen.

[Schultz et al., 2014] Schultz, M., Krabbenhoeft, N., and Skinner, K., editors (2014). *Guidance Documents for Lifecycle Management of ETDs.* Educopia Institute, Atlanta, GA.

[Schöpfel et al., 2016] Schöpfel, J., Prost, H., Malleret, C., Južnič, P., Češarek, A., and Koler-Povh, T. (2016). Dissertations and data. *The Grey Journal*, 12(3):126-148.

[Vompras and Schirrwagen, 2015] Vompras, J. and Schirrwagen, J. (2015). Repository workflow for interlinking research data with grey literature. In *8th Conference on Grey Literature and Repositories*, pages 21-28. National Library of Technology.