Dissertations as Data

Joachim Schöpfel*[1], Eric Kergosien*[2], Stéphane Chaudiron[2], and Bernard Jacquemin[2]

[1]Université Lille Nord de France – ANRT, GERiiCO laboratory – BP 60149, France
[2]Université Lille Nord de France – GERiiCO laboratory – BP 60149, France

Abstract

Problem/goal
The paper provides an overview and empirical evidence on the usability of electronic theses and dissertations (ETDs) and related research data for text and data mining (TDM) techniques.

Research method/procedure

The first part of the paper is a review of recent publications and projects on the potential and usefulness of ETDs for TDM, followed by a description of our own research projects in the field.

Anticipated results

Usually, research studies on dissertations and data address the handling and potential exploitation of dissertations as a "data vehicle", where data are published together with the dissertation (e.g. as a kind of data appendix), or as a "gateway to data", where the data are not published together with the text but are available on a distant server. Yet, often the data are not available; or data, methodology, tools, primary sources are mingled, not indexed, badly described, and unrelated with the text, unconnected with other files.

Our paper will describe a different approach that may be helpful to cope with this problem, in particular (but not only) when it is impossible to distinguish between data and dissertations and thus to process the data appropriately (data repository etc.). Our approach is to consider the dissertation as a whole (text, metadata, data, numbers, facts, figures etc.), as "material" potentially exploitable by TDM tools (including natural language processing) designed for unstructured information, i.e. lacking a pre-defined data model or not organized in a pre-defined manner.

These tools and techniques may be helpful to find patterns or other useful information but usually involve some kind of structuring the documents, e.g. through manual tagging with metadata. A quite different condition is the legal feasibility. While in some countries TDM for scientific purpose does not require copyright clearance because copyright exceptions recognize that it is legal to extract content for data analytics, in other countries like in France copyright-based legal barriers to TDM are still waiting for removal.

Our paper will address these issues, in a general way but also with regards to recent research on content mining of UK dissertations in law and chemistry, to automatic processing

*Speaker

of PhD metadata for innovation search and identification of scientific skills and to our own research projects on TDM of unstructured information in the fields of cultural and industrial heritage, geographical data and academic publishing. In particular, we will draw on preliminary results of our interdisciplinary research project TERRE-ISTEX (2016-2018) that will retrieve, organize and make accessible knowledge related to geographical territories from heterogeneous digital academic resources available on the ISTEX platform and in dissertations.

Also, we will address the issue of retro-digitisation of older print dissertations and related material in order to make them usable for automatic content mining and to valorise these often hidden treasures of academic heritage.

Practical implications/originality

The paper will provide an up-date on an emerging and promising field of research and development. Our results will be useful for academic libraries and repositories, for the conception and creation of added value services for their ETDs.

On the authors:

Joachim Schöpfel is senior lecturer of Library and Information Sciences at the University of Lille (France), Director of the French Digitization Centre for PhD theses (ANRT) and researcher at the GERiiCO Research Center. He was Manager of the INIST (CNRS) scientific library from 1999 to 2008. He teaches Library Marketing, Auditing, Intellectual Property and Information Science. His research interests are scientific information and communication, especially open access, research data and grey literature. He is member of euroCRIS.

Eric Kergosien is senior lecturer of information science at the Department of Information and Document Sciences at the University of Lille and researcher at the GERiiCO Research Center.

Stéphane Chaudiron is a Professor of Information Science at the Department of Information and Document Sciences at the University of Lille and the Director of the GERiiCO Research Center. His main research interests lie in scientific information and communication, information behaviour and practices, evaluation of information retrieval systems, digital humanities and knowledge organization. He is a member of ASIST and ISKO.

Bernard Jacquemin is senior lecturer of information science at the Department of Information and Document Sciences at the University of Lille, head of the Department of Continuous Education and researcher at the GERiiCO Research Center. His research focuses particularly on information structure building, collaborative information design, and practice of digital information services.

Keywords: Electronic theses and dissertations, text and data mining, content mining, retro, digitisation, research data