Dissertations as Data

Joachim Schöpfel, Eric Kergosien, Stéphane Chaudiron, Bernard Jacquemin
GERiiCO laboratory, University of Lille (France)
Overview

• Conceptual approach and empirical application

• About dissertations
• Dissertations and data
• Dissertations as data
  • Examples
  • Barriers
• Concluding remarks
About dissertations

Three ways to describe a dissertation
Dissertation as a document

Source: RTD-DOC, CNRS 2003-2006
Dissertation as a form

= an object whose materiality allows the users to use it and to manipulate it

- Format
- Media
- Structure
- Figures
- References
- Annexes, other material
Dissertation as a sign

= content, a semantic network of information, part of a collection

• Terminology
• Semantics
• Relevance
• Context of interpretation
• Knowledge

bagisme, la dyslipidémie et l’hypertension artérielle.

Parmi les facteurs de risque réversibles, le tabagisme est le plus fortement corrélé avec le risque d’IDM [Yusuf et al. 2004]. Cette corrélation est d’autant plus importante que le sujet est jeune puisque 58% des IDM chez les hommes de moins de 40 ans sont imputables au tabagisme [Cambou 2008]. Le tabagisme passif augmente également le risque cardiovasculaire [Giannuzzi et al. 2005]. Sur le plan physiopathologique, la consommation de cigarettes va entraîner une augmentation du taux de LDL, ce qui induit une réponse inflammatoire qui augmente le risque de rupture de la plaque d’athérome, qui peut conduire à une thrombose [Ambrose et al. 2004]. Par ailleurs, le tabagisme induit une dysfonction endothéliale qui a pour conséquence une diminution du flux sanguin en augmentant l’opposition des artères à ce flux mais également en diminuant la vasodilatation induite par l’oxygène [Ambrose et al. 2004]. L’accumulation de ces différents facteurs va augmenter le risque d’IDM (Figure 9).
Dissertation as a medium

= a trace of social activities, a tangible element of communication between human beings, a social phenomenon

• Institutional character (legitimacy)
  • Prescription (format, content...)
  • Legal and administrative status
  • Process (deposit, preservation, dissemination...)

• Research project (scientificity)
  • Evaluation (impact)
  • Network (altmetrics)
  • Open science (outreach)

• Creative work (authorship)
  • Intellectual property
  • Competitive strategy (career)
  • Financial interests (publication, patent...)
Dissertations and data

A reminder
Data vehicle

- Dissertations contain data
  - Primary data (sources)
  - Secondary data (results)
- In the text
  - Figures
  - Tables
- In appendix
  - Various formats, often non compliant with content mining
  - More or less structured
  - Sometimes on different medium
  - No identifiers
  - Metadata?
Data in PhD dissertations (survey 2015, N=780)
Data in PhD dissertations (survey 2015, N=780)
## Data in PhD dissertations (survey 2015, $N=780$)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Databases</th>
<th>Graphs - figures</th>
<th>Images - drawings</th>
<th>Maps</th>
<th>Others</th>
<th>Photographs</th>
<th>Statistics</th>
<th>Tables</th>
<th>Tests</th>
<th>Timelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archeology</td>
<td>4</td>
<td></td>
<td></td>
<td>22</td>
<td>18</td>
<td>11</td>
<td>1</td>
<td>16</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>Economical sciences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Educational sciences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foreign languages &amp; literatures</td>
<td>1</td>
<td>1</td>
<td></td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>21</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>French language &amp; literature</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geography</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History</td>
<td>16</td>
<td>16</td>
<td>39</td>
<td>27</td>
<td>26</td>
<td>14</td>
<td>14</td>
<td>12</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>History of arts</td>
<td>6</td>
<td>17</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td>14</td>
<td>4</td>
<td>21</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Information science</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Language science</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Law</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management</td>
<td>2</td>
<td>12</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>26</td>
<td>12</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Others</td>
<td>1</td>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Philosophy</td>
<td>2</td>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Political science</td>
<td>1</td>
<td>1</td>
<td></td>
<td>4</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Psychology</td>
<td>2</td>
<td>15</td>
<td>20</td>
<td>1</td>
<td>4</td>
<td>55</td>
<td>65</td>
<td>48</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Sociology</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>21</td>
<td>28</td>
<td>29</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>38</strong></td>
<td><strong>111</strong></td>
<td><strong>175</strong></td>
<td><strong>63</strong></td>
<td><strong>77</strong></td>
<td><strong>65</strong></td>
<td><strong>101</strong></td>
<td><strong>306</strong></td>
<td><strong>376</strong></td>
<td><strong>23</strong></td>
</tr>
</tbody>
</table>

GL17 – Dec 1 2, 2015
Gateway to data

- Dissertations (can) link to data
  - Primary data (sources)
  - Secondary data (results)
- Where is the data?
  - In data repositories
  - In institutional repositories
  - In social networks, ...
- Format
  - Initial (production), preservation, dissemination
  - More or less structured
- Metadata
  - Identifiers (DOI)
  - Generic v disciplinary
Separation of text and data
Workflow – Dataverse ETD Pilot Program

• Metadata
• Documentation
• Identifier
• Link

Source: Dataverse ETD Pilot Program at Emory (Doty et al. 2015)
Workflow – ETDplus project (Educopia Institute)

• Separation of the dissertation text files and the related “complex content objects” whenever possible.

• “Embedding multimedia components within the full text might seem advantageous in that they would then be inseparable. However, when the time comes that it is necessary to migrate either the full text itself or one of the multimedia components, having separate files would greatly simplify matters” (p.5-9).

• Metadata and persistent identifiers like handle, PURL, ARK or DOI are supposed to provide the “glue” that binds together the text, multimedia and data files.

• Preservation-worthy research data (survey data, measurements, laboratory notebooks, measured spectra etc.) might be stored as part of an “ETD package” or, after transformation into archival files, in a separate data repository, and “links to the data repository from the ETD metadata would then enable researchers to access these research data in the future” (p5-11).

Source: Guidance Documents for Lifecycle Management of ETDs (Schultz et al. 2014)
Workflow – Bielefeld PUB

• For working papers but also for ETDs
• Research data as a discrete resource
• Attribution of DOI (DataCite)
• Distinction between institutional and disciplinary data repositories

Source: Bielefeld PUB (Schirrwagen & Vompras 2015)
Dissertations as data

Another way to see and handle dissertations
Data

• No distinction between text and data
• The whole dissertation is data
  • set of values of qualitative or quantitative variables
  • some existing information or knowledge represented in some form suitable for better usage or processing
  • can be measured, collected and reported, and analyzed, to create information suitable for making decisions
• Distinction between data and metadata
  • Yet, metadata is data, too
From print to ...

- Form
  - Medium
  - Inscription

- Sign
  - Content
  - Meaning

- Medium
  - Inscription
  - Legitimacy

Digital

- Form
  - Structure
  - Data

- Sign
  - Structured Data
  - Knowledge

- Medium
  - Data
  - Procedure
ETD as a form

= structure + data; allows for information extraction and filtering

• Format: XML (syntax)
• Media: Web
• Structure: DTD
• Text encoding: TEI (grammar)
• Resource Description Framework
• Metadata

http://etd.vt.edu/etd-ml/dtdetds.htm
ETD as a **sign**

= structured data + knowledge; aboutness, intentionality

- Semantics
- Ontologies
- Vocabulary
- Geolocation
- Research information

ETD as a medium

= data + procedure

• Rules (prescription)
  • Format
  • Structure

• Legal status
  • IP exception
  • Licensing
  • Access restrictions

https://wikis.tdl.org/tdl/ETD_Workflow_Diagram
Re-use to enable new insights from text and data

• Text and data mining (TDM) or content mining of machine-readable text and data
• Extraction of information (knowledge) from large volume of content
• “The main barriers against the uptake of TDM are not technical, but primarily a lack of awareness among academics, and a skills gap. They relate to legal issues around copyright and database rights, and to some policy choices of restrictions being implemented by publishers on (for instance) access to APIs. These problems are all soluble, but require non-technical solutions.”

SPECTRa-T – a JISC project 2007-2008

• A proof-of-concept approach focusing on chemistry research data in molecular and related subjects

• To develop text-mining tools and processes for the automatic extraction of experimental research data (chemical objects and named chemical entities)

• To transform the extracted data into metadata and ingest them into data repositories and RDF triplestores, thus enabling RDF-based semantic querying of the contents

SPECTRa-T workflow

Various ETD formats – LaTeX, Postscript, Adobe Portable Document Format (PDF), Microsoft Word (DOC), Microsoft Office Open XML (DOCX)

PDF flow

THESIS
Input document (text)

SPECTRa-T
Text processing tools

OSCAR3

XSL stylesheet

SciXML

SAFXML

RDF

Query

Triplestore

http://www.dspace.cam.ac.uk/handle/1810/230116
SPECTRa-T on ETDs as form

• “PDF (…) is significantly less well-suited to text-mining than is marked-up text.”

• “For data mining in chemistry, a marked-up document file format (such as Office Open XML) for deposited e-theses should be available in preference to PDF.”

• “Require deposit of the Word version of the thesis, and then to convert it to XML as a standard process (…)”
SPECTRa-T on ETDs as

- “Discipline-specific text-mining tools, together with the triplestore and data repositories into which text-mined outputs are to be deposited, will best be managed by subject experts operating within the same disciplinary environment.”

- “Much of the effort involved in developing text-mining tools and workflows is unavoidably specific to individual disciplines.”

- “Those who are responsible for designing and delivering institutional repository services need to work closely with researchers in different subjects, in order to understand their needs and co-ordinate departmental and institutional activities.”
“Institutions to determine what purpose theses might serve in each discipline beyond their immediate role in assessing a student's research (…)"

“In the sciences theses should be regarded not simply as objects destined for preservation and archival status, but rather as unique resources containing potentially valuable data that must be made extractable and re-usable.”

“Those authorities with responsibility for overseeing the writing, submission and subsequent management of theses (should) implement appropriate strategies, applying both institution-wide and discipline-specific policies, guidelines and practices designed to ensure optimal conditions for re-use. In other words, institutional regulations should ensure that each thesis is fit for purpose.”
TERRE-ISTEX project for geographic indexing and analysis from scientific corpora

Thematic project (funded by ISTEX 2016 – 2017)
• 1 objective: To offer an unique and exceptional concentration of scientific documentation archives
• 2 parts: acquisition of scientific digital archives, creation of the ISTEX platform
• 3 level of services: basic services, value added services, thematic projects
• 5 partners (+): CNRS, ABES, Couperin, MESR, Université de Lorraine for engineering CPU (+ R&D)
• Budget: 60m € (« Investissements d’Avenir »)
• Duration: 3 years
• http://www.istex.fr/
• To study what happens on a territory on the thematic « climate change » from heterogeneous scientific data (scientific articles from CIRAD and ISTEX, PHD Thesis)
• To give an help to Scientists to do their state-of-the-art (on a territory, on a thematic...)

Questions :
• What is a **territory**? → A set of geographical information
  - Geographical information = spatial entity (where) + thematic entity (what) + temporal entity (when)
  
  → Example : a study on climate changes realized in the south of Madagascar in 1981.
• How to identify this kind of information? → text mining methods in abstract and content, experts validation

A use case for research information:
• What are the scientific documents related to Madagascar and Sénégal?
• In these documents, what are the studied thematics and the associated authors?
  → Development of a Geographical Research Tool based on Elastic Search
  → Spatial visualization
  → the results analysis is made by experts
CIRAD data

Multilingual data Données coming from the Agrotrop database: CIRAD open archives
92,000 references and 25,000 complete documents: scientific publications and grey literature (reports, etc.)
3,000 documents related to the thematic «climate change»
Metadata: title, author, abstract, indexed labels coming from lexicons (AGROVOC thesaurus and Agris/Caris from the FAO). Thèmes Agris: https://agritrop.cirad.fr/view/subjects/, geographical information (countries)

PHD Thesis data (ANRT)
200,000 PhD theses, metadata, SUDOC link (ABES).
70,000 digital documents
ABES: RAMEAU thesaurus from the BnF for label description
400 PhD theses related to the thematic «climate changes»

ISTEX Data
8,190,000 scientific articles (Emerald, Wiley, Elsevier, etc.) related to the thematic «climate change» (10,000 are in french)
Metadata into MODS (XML format)
Lille: General process

Identification of relevant data
Content and meta-data:
- Places and spatial geocodes
- Year of publication
- Themes and/or domains
- Abstracts

Data validation

Geographic analysis
- Space
- Time
- Thema

Information retrieval
- Thema, Time, Space, Full-text

Documents
Series of conferences
And PhD thesis

Indexation
Geographical analysis on the EGC conference publications (1200 articles): Indexing, Research information, Analysis

• Build a first geographic index (MODS model for metadata)

<table>
<thead>
<tr>
<th>Information thématique</th>
<th>Nom ville conférence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noms villes auteurs</td>
</tr>
<tr>
<td></td>
<td>Noms auteurs</td>
</tr>
<tr>
<td></td>
<td>Titre article</td>
</tr>
<tr>
<td></td>
<td>Résumé article</td>
</tr>
<tr>
<td></td>
<td>Session</td>
</tr>
<tr>
<td></td>
<td>Domaine</td>
</tr>
<tr>
<td>Information spatiale</td>
<td>Coordonnées villes auteurs</td>
</tr>
<tr>
<td></td>
<td>Coordonnées ville conférence</td>
</tr>
<tr>
<td>Information temporelle</td>
<td>Année conférence</td>
</tr>
<tr>
<td>Information plein texte</td>
<td>Termes titre article</td>
</tr>
<tr>
<td></td>
<td>Termes résumé article</td>
</tr>
</tbody>
</table>

Implementation of a Multicriteria Research Information tool based on Elastic Search
The sentiAnnotator viewer for the extraction process and experts validation (ISWC, October 2015)

- Upload of Big corpora (french and english)
- NLP process to extract features
- Experts can (un)validate marked data
- Possibility to download results
Lille: First work (proof of concepts)

• Geographical analysis on the EGC conference publications: Indexing, Research information, Analysis

• A set of geographical analysis available (in french language): Elastic Search and Kibana tool
  http://ekergosien.net/DefiEGC/index.html
Lille : First work (proof of concepts)

- A set of geographical analysis available: Elastic Search and Kibana tool
A set of geographical analysis available
Prospects: « big data », Valorization

• « Big data »
  • Heterogeneous data from ISTEX, PhD theses and Cirad corpus :
    → Normalization of metadata with the MODS model (XML)

• Volume:
  → Test the scalability of our text mining application on bigger corpora

• Tests on other thematics (history, technology, culture...)

• Integrate other resources (web, bases de données, archives ouvertes, plateformes...)

ETD2016 Lille 11-13 July 2016
Master and BA dissertations

• Indexing students dissertations (for Lille 3 students: thousands of dissertations)
• Implementation of a dynamical geographical information research tool
Barriers
Availability

• One part of dissertations is not digital

• Especially older dissertations

• Digitization of print dissertations: problem with OCR (see SPECTRa-T)

Accessibility

- One part of dissertations is not online
  - And/or not (easily) identifiable
- Some (few) are confidential
- More are embargoed (publishing)
- Others are restricted to on-campus access only
- Social networks, private websites...

http://www.dlib.org/dlib/march15/schopfel/03schopfel.html
Legality

• Dissertations are online but TDM is not allowed
  • No law
  • No license

• Gratis v libre
  • Harnad v Murray-Rust

• Orphan dissertations?

18. ALLOW TEXT AND DATA MINING

“Text & data mining” (TDM) or “content mining” refers to a variety of analytical procedures to derive additional information and knowledge from a large set of publications.

EXAMPLE: A scientist may use a computer to automatically analyse a large number of research papers to discover a previously unknown association between two diseases.

Some rightsholders insist that this requires two separate licences, one for reading the works and another one for TDM.

A specific and mandatory exception to remove text and data mining for scientific purposes from the reach of European copyright and database law should be considered. — EC TDM expert group recommendation

THE REPORT RECOMMENDS: Clarify that lawful access to data includes the right to mine it through automated analytical techniques.

https://juliareda.eu/copyright-evaluation-report-explained/#tdm
Feasability

- Dissertations are online (ETDs) but TDM is not possible (optimal)
- Format (PDF processing not optimal for TDM)
- Structure (non recognition of relevant sections)
- Metadata (discipline, domain...)
- Discipline-specific knowledge (and tools)

Concluding remarks
Open science

• Amsterdam Call for Action (2016)
  • EU copyright reform and national initiatives
  • Licensing and TDM-friendly formats by publishers
• “The Right to Read is the Right to Mine” LERU 2016 (League of European Research Universities)
• EU copyright reform (2017?), objective TDM:
  • “A mandatory exception for research and education purposes;
  • a mandatory exception that will enable users to text and data mine all content to which they have legal access”
• French new Law for a Digital Republic (2016):
  • TDM exception for scientific documents (if legally accessed)
  • For scientific purpose, no commercial use
Conditions – A2LF

- Availability
- Accessibility
- Legality
- Feasibility
## Recommandations

### Availability
- Retrodigitization
- Text and complementary material

### Accessibility
- Libre open access
- Institutional repositories

### Legality
- TDM exception (or fair use)
- Clearance of third party rights (prescription rules)

### Feasibility
- Standards (prescription rules)
- Rich metadata and file structure
References


• All references available at http://www.citeulike.org/user/Schopfel/tag/etd2016

• Corresponding author: joachim.schopfel@univ-lille3.fr
That's all Folks!