



Research data in social sciences and humanities

Joachim Schöpfel, University of Lille



RESEARCH DATA IN SS&H



Contents

- ❖ Terminology and categories
- ❖ Data and publications
- ❖ Critical issues

- ❖ NOT Research data management



Terminology and categories



TERMINOLOGY



Data are like cows. If you look them in the face hard enough they generally run away.

(adapted from Dorothy L Sayers)

« Recorded factual material commonly accepted in the scientific community as necessary to validate research findings »

US OMB Circular 110

« Re-usable research results, collected, observed or created for purposes of analysis to produce original results »

University of Edinburgh

TERMINOLOGY



Definition mainly by functions (validation, reuse, innovation) and types (and not by nature)

- “Research data refers to information, in particular facts or numbers,
- collected to be examined and considered and as a basis for reasoning, discussion, or calculation.
- In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images”

H2020



October 24, 2016

What is information? Numbers = data?

But what are facts?



GENERAL TYPOLOGY



- ❖ Research methods
 - ❖ Observational data
 - ❖ Experimental data
 - ❖ Simulation data
 - ❖ Derived or compiled data
- ❖ Input and output
 - ❖ Primary data (collected)
 - ❖ Secondary data (produced)



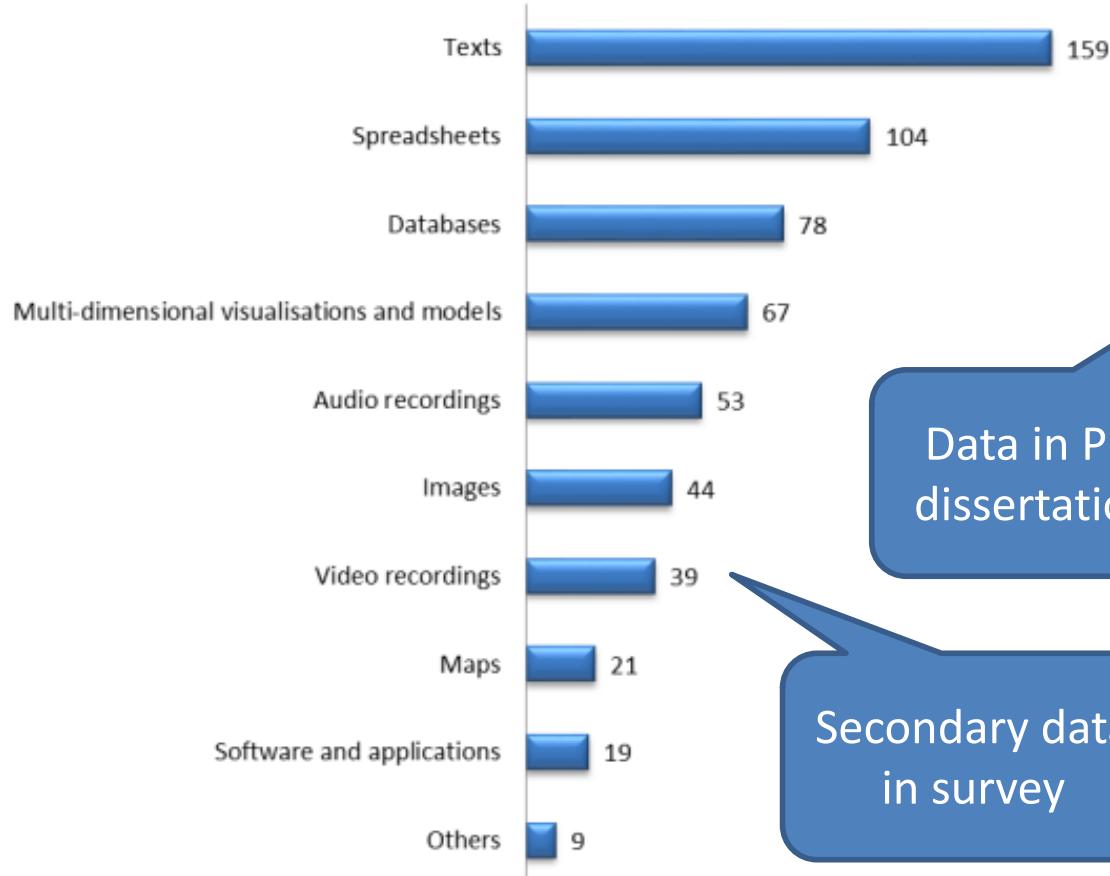
CATEGORIES



- Archived data
 - Audiovisual data
 - Configuration data
 - Databases
 - Images
 - Networkbased data
 - Plain text
 - Raw data
 - Scientific and statistical data formats
 - Software applications
 - Source code
 - Standard office documents
 - Structured graphics
 - Structured text
 - other
- Observations
 - Experiments
 - Simulations
 - Images
 - Surveys and interviews
 - Statistics and reference data
 - Logfiles and usage data
 - Text documents
 - Other (please specify)

re3data, HUB

CATEGORIES IN SS&H



- Texts
- Tables
- Images, drawings
- Graphs, figures
- Statistics
- Maps
- Photographs
- Databases
- Timelines
- Others

Data in PhD
dissertations

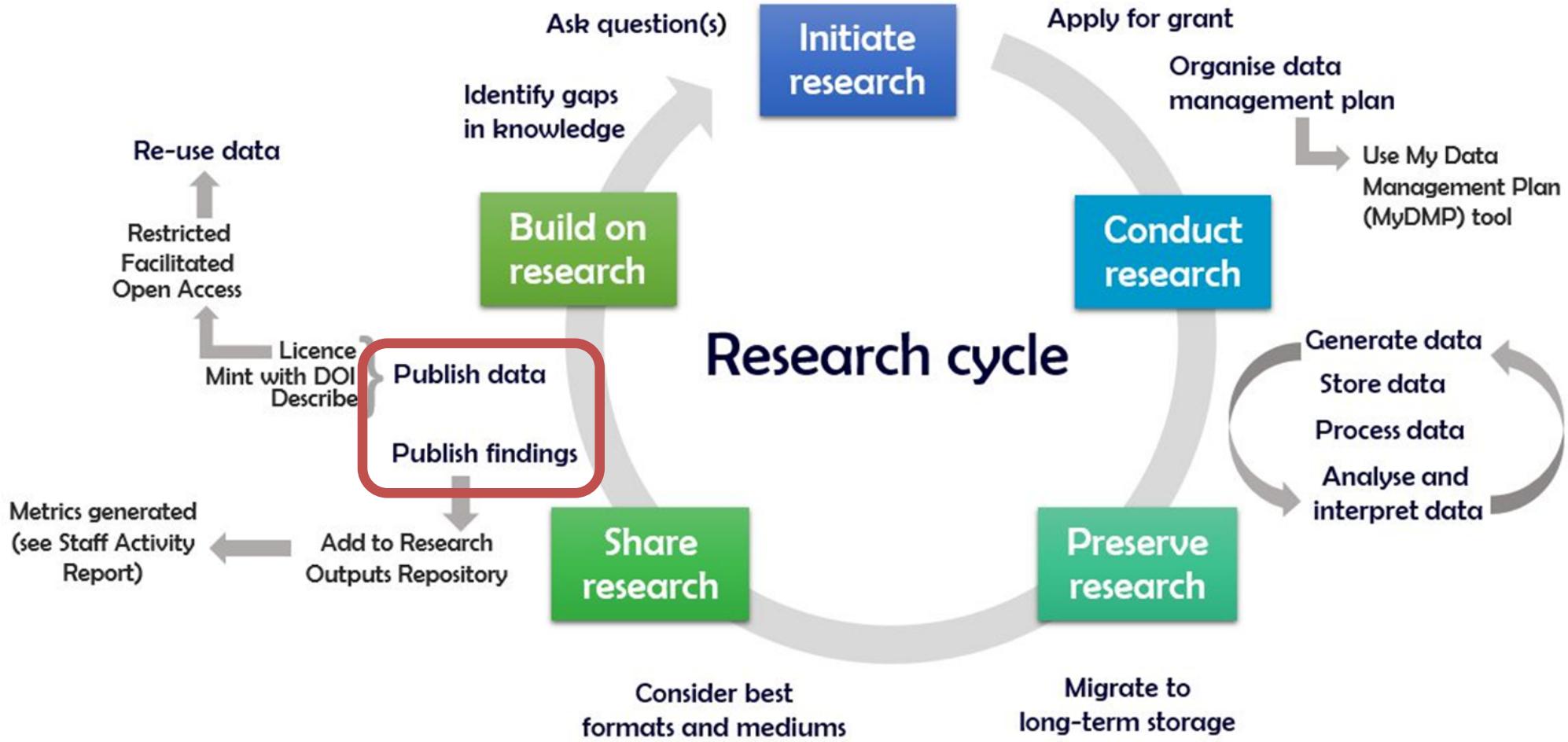
Secondary data
in survey

Lille studies



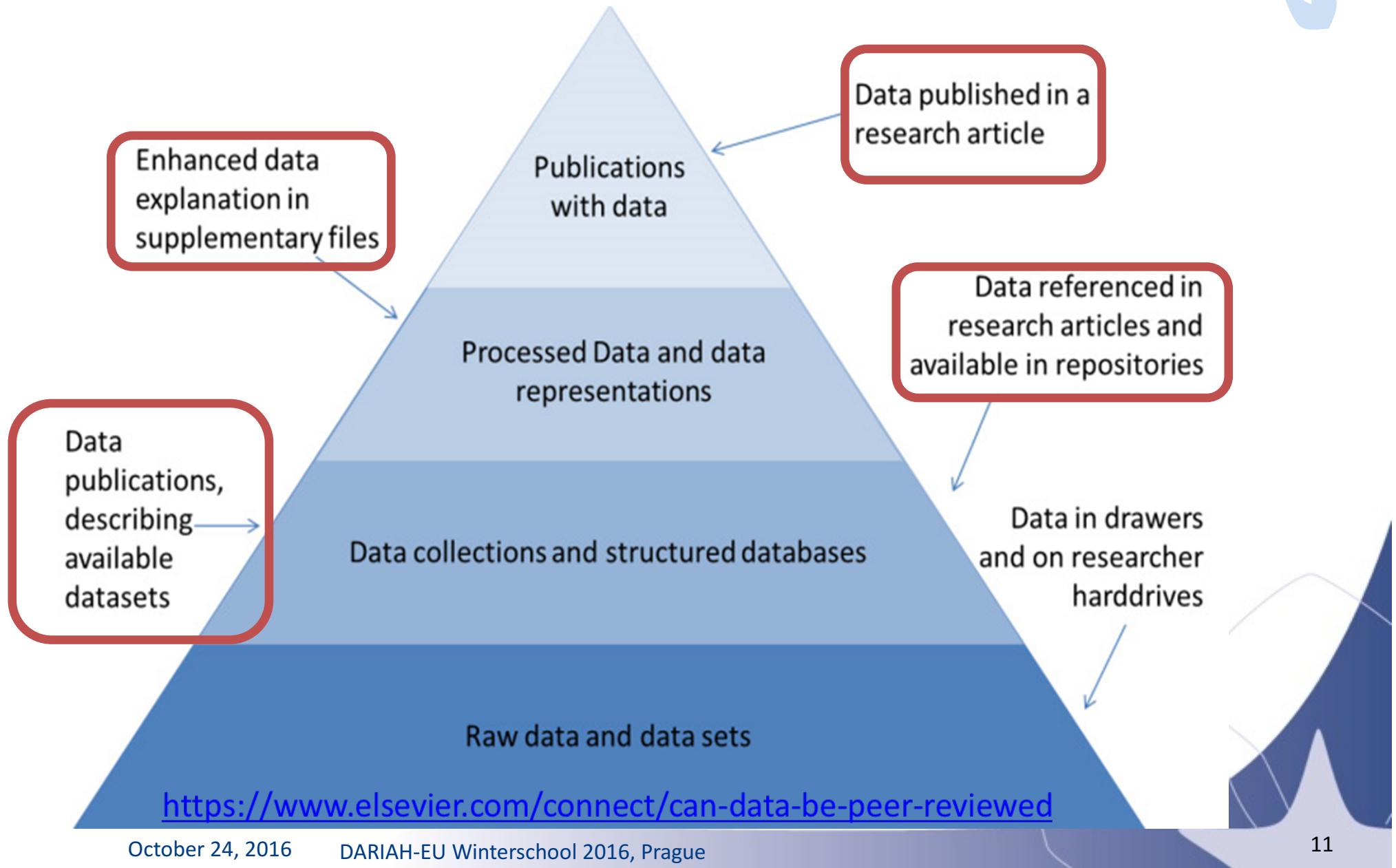
Data and publications

RESEARCH AND DATA



Source: <http://guides.library.unisa.edu.au/ResearchDataManagement>

DATA PUBLICATION



DATA AND PUBLICATIONS



- ❖ Data vehicle
 - ❖ Supplementary materials of publication
- ❖ Document as data
 - ❖ Exploited as primary data source for TDM
- ❖ Gateway to data
 - ❖ Publication contains links to data, integrated or not in the text

EXAMPLE DISSERTATIONS



ETD

= 8 =
Athènes, où l'industrie du bâti était fort en honneur, les plaideurs de la troupe de Pégase étaient rares, et si Solon avait singé ce débat, le préteur aurait bien été fâché de s'en occuper.
Malheureusement il y avait des cas où l'observation de la loi était impossible en fait ou en droit. Il s'agissait d'un enfant enlevé, d'un absent ou d'une femme, il était nécessaire que quelqu'un pût en moins laisser la cause au plaigneur, qui ne pouvait se servir pas pourtant lui-même sa défense. Si l'État était directement intéressé à une affaire, si par exemple il s'agissait d'un procès de haute trahison, il ne lâcherait pas qu'il n'y eût assez grave en déshonneur de l'intérêt public. En abandonnant à l'initiative privée, la république désignait donc des créatures, chargée de la représentation, pour assurer l'assistance. Dans tous ces cas, il avait bien fallu déroger à la règle, et admettre que le plaigneur devait faire preuve de patience et de persévérance. D'autre part, dans certains cas, il était tout à fait nécessaire de faire preuve de patience et de persévérance pour une recouvrement. Tanguy qui devait assurer à leur assistance des fonctionnaires démissionnaires. Il avait pour partie que ce plaigneur incomptemps puis appeler à son aide pour l'assister et discuter, dans la logique de l'Etat, pour assurer à ce démissionnaire de faire son travail, mais il avait bien fallu s'adonner à quelques manœuvres, dont on retrouve dans l'ordre de l'Etat, et d'un autre.

Data

ETD



ETD

= 8 =
Athènes, où l'industrie du bâti était fort en honneur, les plaideurs de la troupe de Pégase étaient rares, et si Solon avait singé ce débat, le préteur aurait bien été fâché de s'en occuper.
Malheureusement il y avait des cas où l'observation de la loi était impossible en fait ou en droit. Il s'agissait d'un enfant enlevé, d'un absent ou d'une femme, il était nécessaire que quelqu'un pût en moins laisser la cause au plaigneur, qui ne pouvait se servir pas pourtant lui-même sa défense. Si l'État était directement intéressé à une affaire, si par exemple il s'agissait d'un procès de haute trahison, il ne lâcherait pas qu'il n'y eût assez grave en déshonneur de l'intérêt public. En abandonnant à l'initiative privée, la république désignait donc des personnes, chargée de la représentation, et de assurer l'assistance. Dans tous ces cas, il avait bien fallu déroger à la règle, et admettre que le plaigneur se pouvait faire représenter par autres.
De même, il y a également d'un bâti, d'un bâtiment, d'un modèle ou d'un modèle, dans la langue indienne un monstre rebelle à la peinture, il devient alors difficile d'appliquer la loi. Cela est vraiment possible lorsque les peintres jettent un peu de peinture sur le bâti, qui dans ce cas empêche d'exprimer le premier tout de son affiche. N'importe pas l'indien une question certaine que de peintres viennent bâti bâti avec le bâti, qui dans ce cas empêche d'exprimer le deuxième d'un plaigneur court, devant expérimenter sous forme de démissionnaires. Tanguy qui devait assurer à leur assistance des fonctionnaires démissionnaires. Il avait pour partie que ce plaigneur incomptemps puis appeler à son aide pour l'assister et discuter, dans la logique de l'Etat, pour assurer à ce démissionnaire de faire son travail, mais il avait bien fallu s'adonner à quelques manœuvres, dont on retrouve dans l'ordre de l'Etat, et d'un autre.

Data

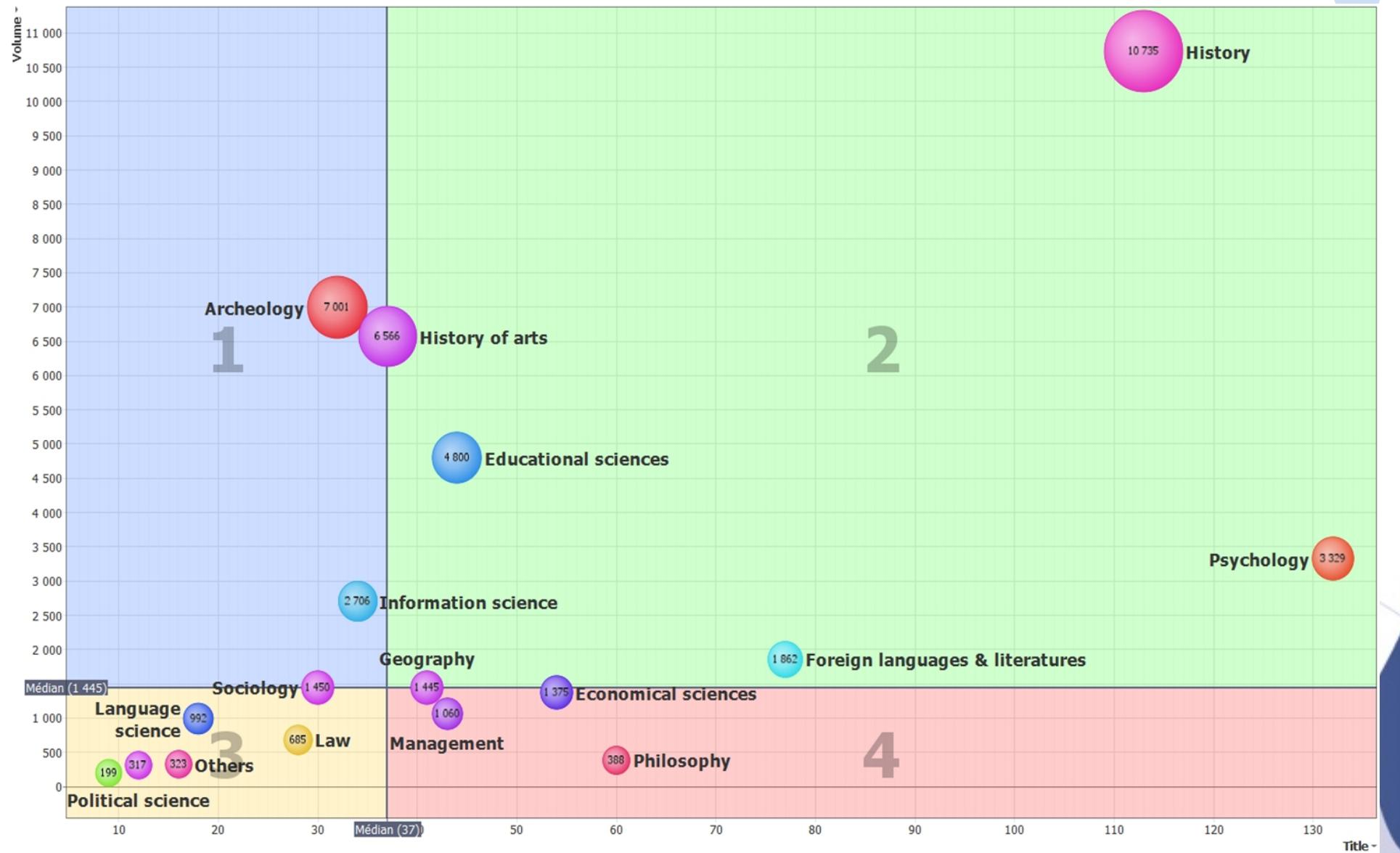
Data

ETD as data vehicle

ETD as data

ETD as gateway to data

DATA IN DISSERTATIONS



DATA IN DISSERTATIONS CT'D



Y: Domain	Databases	Graphs - figures	Images - drawings	Maps	Others	Photographs	Statistics	Tables	Texts	Timelines	Tous
Archeology	4	2	22	18		11	1	16	15	1	30
Economical sciences		16	1	5			2	31	36		43
Educational sciences		8	14	1			5	25	29	1	38
Foreign languages & literatures	1	1	20		1	1	6	21	36	1	46
French language & literature		1					1		5	1	6
Geography		13	7	13		5	3	27	23		33
History	16	22	39	27		26	14	44	65	12	88
History of arts	6		17	8	1	8		4	20	1	28
Information science	2	7	7	3	4	2	5	12	20	1	28
Language science	1	1	1				1	1	7		7
Law		1	3	2				4	5		7
Management	2	12	10	1		1	7	26	22	2	30
Others	1	2						2	4	1	6
Philosophy		2	2		1	1		1	11		11
Political science	1	1	4				1	6	2		6
Psychology	2	15	20	1		4	55	65	48		91
Sociology	2	7	8	4		6		21	28	2	28
Tous	38	111	175	83	7	65	101	306	376	23	526

X: Typology

DATA IN DISSERTATONS - ISSUES



Incomplete,
inadequate
or missing
description

Missing
organisation

Inadequate
format

Datasets and/or
individual data are
not or incompletely
documented

Research data are
presented without
any structuration or
organisation, often
together with other,
not reusable
material in a kind of
information mash-up
not suitable for
further research

Data and text are
glued together in a
PDF file instead of
being separated and
published in
adequate file
formats

ENHANCED ARTICLE



Palgrave communications

❖ Data available

❖ (not found)

❖ Data not available

❖ Confidentiality/privacy issues

❖ Or only on demand

❖ No data

❖ “Data sharing is not applicable to this article, as no datasets were generated or analysed during the current study.”

RESEARCH REPORT



Université
de Lille
3 SCIENCES HUMAINES
ET SOCIALES

Publications scientifiques de l'Université de Lille Sciences Humaines et Sociales



Accueil

Dépôt

Consultation ▾

Aide ▾

hal-01198379, version 1

Les données de la recherche en SHS. Une enquête à l'Université de Lille 3. : Rapport final

Hélène Prost^{1, 2}, Joachim Schöpfel² [Détails](#)

¹ INIST - Institut de l'information scientifique et technique

² GERICO - Groupe d'Etudes et de Recherche Interdisciplinaire en Information et COnnexion

Résumé : Les données de la recherche deviennent l'un des nouveaux défis de la gestion scientifique. Produites dans le cadre des projets de recherche, ces données posent des questions inédites aux laboratoires, bibliothèques et services informatiques des universités : comment conserver ces données, comment les signaler et mettre à disposition d'autres chercheurs, comment faire le lien avec les publications, comment les intégrer dans une politique du libre accès à l'information scientifique ? Avant de se lancer dans un projet de données de la recherche, un établissement doit faire un état des lieux sur le terrain pour mieux connaître les producteurs de ces données, leurs pratiques et besoins dans la gestion des données et leurs outils, mais aussi la nature concrète de ces données. L'Université de Lille 3 a réalisé une étude sur les pratiques, besoins et attentes en matière de gestion des données de la recherche auprès de son personnel scientifique. L'étude est pilotée par le laboratoire GERICO et le SCD de Lille 3. Elle fait partie d'une démarche concertée en faveur de la gestion et du partage des données de la recherche mise en œuvre à partir de 2013, avec plusieurs analyses, séminaires et publications. L'enquête a été préparée avec l'Université Humboldt de Berlin. Le questionnaire contient 22 questions et a été mis en ligne en avril et mai 2015. Il a reçu 270 réponses (taux de réponse 15%). Toutes les disciplines sont représentées, ainsi que toutes les catégories des personnels scientifiques. Les personnes interrogées décrivent un large éventail de données sources. Les corpus (documents textuels) sont de loin la source la plus importante (64%), suivie par les enquêtes et entretiens (47%), observations (41%), expériences (36%) et archives (34%). Quant à la typologie des données

FICHIERS



Rapport enquête données de l...
Fichiers produits par l'(les)
auteur(s)

Autre - *Commentaire : Fichier Excel regroupant les réponses récoltées suite à l'enquête*

Masquer les fichiers annexes

LICENCE



Distributed under a Creative Commons Paternité 4.0 International License

IDENTIFIANTS

- HAL Id : [hal-01198379](http://hal.univ-lille3.fr/hal-01198379), version 1



<http://hal.univ-lille3.fr/hal-01198379v1>

COMMUNICATION (1)



Publications scientifiques de l'Université de Lille Sciences Humaines et Sociales

Accueil Dépôt Consultation ▾ Aide ▾

hal-01285304, version 1

Dissertations and Data : keynote address

Joachim Schöpfel¹, Južnič Primož², Hélène Prost¹, Cécile Malleret³, Ana Češarek², Teja Koler-Povh² Détails

¹ GERIICO - Groupe d'Etudes et de Recherche Interdisciplinaire en Information et Communication

² University of Ljubljana (SLOVENIA)

³ SCD - Service Commun de la Documentation

Abstract : The keynote provides an overview on the field of research data produced by PhD students, in the context of open science, open access to research results, e-Science and the handling of electronic theses and dissertations. The keynote includes recent empirical results and recommendations for good practice and further research. In particular, the paper is based on an assessment of 864 print and electronic dissertations in sciences, social sciences and humanities from the Universities of Lille (France) and Ljubljana (Slovenia), submitted between 1987 and 2015, and on a survey on data management with 270 scientists in social sciences and humanities of the University of Lille 3. The keynote starts with an introduction into data-driven science, data life cycle and data publishing. It then moves on to research data management by PhD students, their practice, their needs and their willingness to disseminate and share their data. After this qualitative analysis of information behaviour, we present the results of a quantitative assessment of research data produced and submitted with dissertations. Special attention is paid to the size of the research data in appendices, to their presentation and link to the text, to their sources and typology, and to their potential for further research. The discussion puts the focus on legal aspects (database protection, intellectual property, privacy, third-party rights) and other barriers to data sharing, reuse and dissemination through open access. Another part adds insight into the potential handling of these data, in the framework of the French and Slovenian dissertation infrastructures. What could be done to valorise these data in a centralized system for electronic theses and dissertations (ETDs)? The topics are formats, metadata (including attribution

FICHIER



GL17 DissData keynote paper 5 ...
Fichiers produits par l'(les)
auteur(s)

LICENCE



Distributed under a Creative Commons Paternité 4.0
International License

IDENTIFIANTS

- HAL Id : hal-01285304, version 1

COLLECTIONS

UNIV-LILLE3 | GERIICO



<http://hal.univ-lille3.fr/hal-01285304>

COMMUNICATION (2)



Titre du congrès	GL17 International Conference on Grey Literature
Audience	Internationale
Invité	Non
Comité de lecture	Oui
Acte	Non
Vulgarisation	Non
Langue du document	anglais
Pays	Pays-Bas
Domaine	Sciences de l'Homme et Société/Sciences de l'information et de la communication
Voir aussi	http://dx.doi.org/10.17026/dans-xg6-xnj4
Mots-clés	en research data, data repository, institutional repository, open access, open data, electronic theses and dissertations, research data management, Open science

Masquer la liste complète des métadonnées

<http://hal.univ-lille3.fr/hal-01285304>

Contributeur : Laboratoire Geriico <lab.geriico@gmail.com>

Soumis le : mercredi 9 mars 2016 - 18:53:06

Dernière modification le : mercredi 30 mars 2016 - 00:21:39

Document(s) archivé(s) le : lundi 13 juin 2016 - 08:46:10

COMMUNICATION (3)



HOME REGISTER LOG IN



EASY

Get exposure and credit for your data: write a data paper for the new peer reviewed, online-only open access Research Data Journal (published by Brill)

For more info: brill.com/rdj

EASY offers sustainable archiving of research data and access to thousands of datasets.

Search...

SEARCH

» Search help

» Advanced search » Browse

DISSERTATIONS AND DATA

Overview Description Data files (3)

Cite as:

Schöpfel, Dr. J. (University of Lille 3) (2015): *Dissertations and Data*. DANS. <http://dx.doi.org/10.17026/dans-xg6-xnj4>

2015-12-02 | Schöpfel, Dr. J. (University of Lille 3) | [10.17026/dans-xg6-xnj4](http://dx.doi.org/10.17026/dans-xg6-xnj4)

We present the results of a quantitative assessment of research data produced and submitted with dissertations. Special attention is paid to the size of the research data in appendices, to their presentation and link to the text, to their sources and typology, and to their potential for further research. The discussion puts the focus on legal aspects (database protection, intellectual property, privacy, third-party rights) and other barriers to data sharing, reuse and dissemination through open access.

Another part adds insight into the potential handling of these data, in the framework of the French and Slovenian dissertation infrastructures. What could be done to valorize these data in a centralized system for electronic theses and dissertations (ETDs)? The topics are formats, metadata (including attribution of unique identifiers), submission/deposit, long-term preservation and dissemination. This part will also draw on experiences from other campuses and make use of results from surveys on data management at the University of Lille 3-LU

<https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:64209>

Relations

WORKING PAPER



EconPapers

Economics at your fingertips

[EconPapers Home](#)

[About EconPapers](#)

[Working Papers](#)

[Journal Articles](#)

[Books and Chapters](#)

[Software Components](#)

[Authors](#)

[JEL codes](#)

[New Economics](#)

[Papers](#)

[Advanced Search](#)

[Quick Search](#)

[EconPapers FAQ](#)

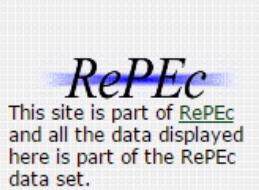
[Archive maintainers FAQ](#)

[Cookies at EconPapers](#)

[Format for printing](#)

[The RePEc blog](#)

[The RePEc plagiarism page](#)



RePEc

This site is part of RePEc
and all the data displayed
here is part of the RePEc
data set.

The Geometry of Finite Equilibrium Datasets

Yves Balasko (yves@balasko.com) and Mich Tvede

No 09-07, [Discussion Papers](#) from [University of Copenhagen, Department of Economics](#)

Abstract: We investigate the geometry of finite datasets defined by equilibrium prices, income distributions, and total resources. We show that the equilibrium condition imposes no restrictions if total resources are collinear, a property that is robust to small perturbations. We also show that the set of equilibrium datasets is pathconnected when the equilibrium condition does impose restrictions on datasets, as for example when total resources are widely non collinear.

Keywords: [equilibrium manifold](#); [rationalizability](#); [pathconnectedness](#) (search for similar items in EconPapers)

JEL-codes: [D31](#) [D51](#) (search for similar items in EconPapers)

Date: 2009-03

References: [View references in EconPapers](#) [View complete reference list from CitEc](#)

Citations [View citations in EconPapers](#) (1) [Track citations by RSS feed](#)

Downloads: (external link)

http://www.econ.ku.dk/english/research/publications/wp/dp_2009/0907.pdf/ (application/pdf)

Related works:

Journal Article: [The geometry of finite equilibrium datasets](#) (2009) [downloads](#)

This item may be available elsewhere in EconPapers: [Search](#) for items with the same title.

Export reference: [BibTeX](#) [RIS](#) (EndNote, ProCite, RefMan) [HTML/Text](#)

Persistent link: <http://EconPapers.repec.org/RePEc:kud:kuiedp:0907>

[Access Statistics](#) for this paper

[More papers](#) in Discussion Papers from [University of Copenhagen, Department of Economics](#) Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark. Contact information at [EDIRC](#).
Series data maintained by Thomas Hoffmann (thho@kb.dk).



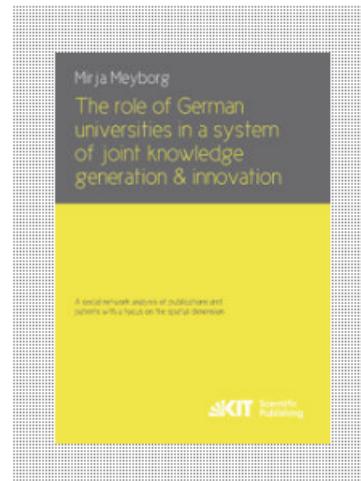
<http://econpapers.repec.org/paper/kudkuiedp/0907.htm>

OTHER PUBLICATIONS



THE ROLE OF GERMAN UNIVERSITIES IN A SYSTEM OF JOINT KNOWLEDGE GENERATION AND INNOVATION | Mirja Meyborg

- ❖ Books – OpenEdition
- ❖ Data journals (later)



14. Appendix

p. 251-271

TEXT ANMERKUNGEN ABBILDUNGEN

VOLLTEXT



Appendix 1: Overview of all German universities that are subject of study (own illustration).

Overview of all German Universities that are Subject of Study									
German Universities	Actors-Code	Elite University	Medical University	Technical University	German Universities	Actors-Code	Elite University	Medical University	Technical University
Aachen RWTH	11	x	x	x	Hildesheim	65			
Augsburg	19				Hohenheim	72			
Bamberg	20				Ilmenau	86			x
Bayreuth	21		x		Jena	87		x	x
Berlin FU	75	x	x		Kaiserslautern	38		x	x
Berlin HU	76	x	x		Kassel	39			
Berlin TU	77			x	Kiel	40		x	
Bielefeld	22		x		KIT	10	x		x
Bochum	23		x		Koblenz Landau	41		x	
Bonn	24				Köln	43	x	x	
Braunschweig	25			x	Konstanz	17	x		
Bremen	26	x	x		Leipzig	88		x	
Chemnitz	78			x	Lübeck	67		x	
Clausthal	61			x	Lüneburg	68			
Cottbus	79			x	Magdeburg	89		x	
Darmstadt	28			x	Mainz	44		x	
	--					--			

<http://books.openedition.org/ksp/244>

PUBLICATIONS AS DATA



- ❖ TDM of research publications
 - ❖ The case of ETDs
 - ❖ The future: writing dissertations as usual?
- ❖ Legal issues
- ❖ Technical issues
- ❖ Impact on publications?
 - ❖ Structure
 - ❖ Content
 - ❖ Format



Critical issues

GENERAL ISSUES



- ❖ Separation text/data
- ❖ Preferred formats cf. [DANS](#)
- ❖ Metadata (generic, specific)
- ❖ Persistent identifiers (DOI, ORCID)
- ❖ Altmetrics (DOI) and usage (low?)
- ❖ Backup, storage, preservation, sharing, reusage
 - ❖ Quality of data repositories (Data Seal of Approval...)

ON DISCIPLINARITY

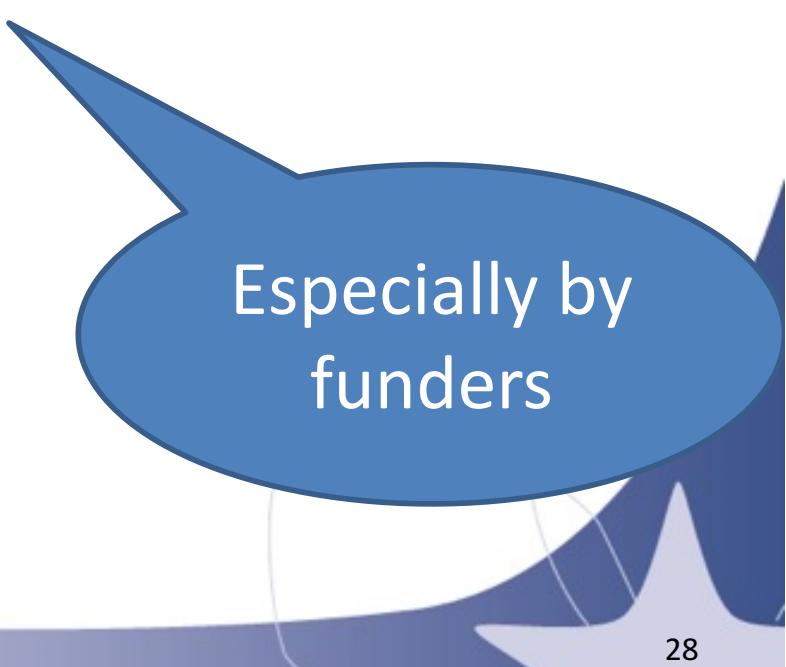


- ❖ Impact of disciplines
 - ❖ Typology
 - ❖ Or rather, profiles
 - ❖ Related to methodology and instruments (ex survey data...)
- ❖ Evaluation: a strong need for a standard and generic approach
 - ❖ Metadata
 - ❖ Identifiers (DOI, handle...)
- ❖ Preservation and sharing: disciplinary and generic repositories
 - ❖ Cf. re3data.org
 - ❖ Cf. HAL, Figshare

RESEARCH EVALUATION



- ❖ Research data are evaluated as research output
- ❖ Mix between primary and secondary data
- ❖ Contrary to publication, data quality or volume are NOT evaluated...
- ❖ ...but the data management
 - Existence of a DMP
 - Description and identification
 - Conservation
 - Sharing



Especially by
funders

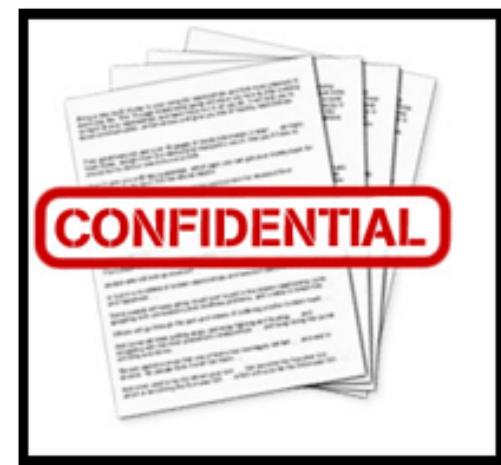
LEGAL ISSUES



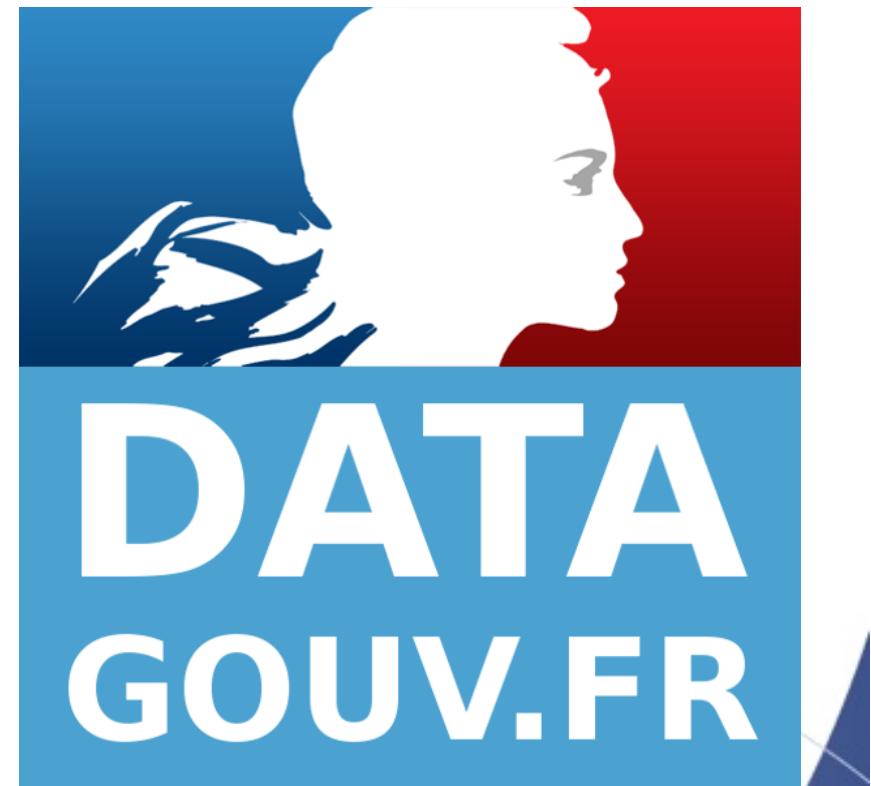
- ❖ Intellectual property?
 - ❖ Career strategy
 - ❖ Publications
- ❖ Data base protection (sui generis)?
- ❖ Third party rights?



- ❖ Confidentiality?
 - ❖ Private company information
 - ❖ Corporate secrets
- ❖ Privacy issues?



POLITICAL ISSUES (1)



POLITICAL ISSUES (2)



Netherlands' EU presidency on 4 and 5 April 2016. It is a living document reflecting the present state of open science evolution. Based on the input of all participating experts and stakeholders' as well as outcomes of preceding international meetings and reports, a multi-actor approach was formulated to reach two important pan-European goals for 2020:

1. Full open access for all scientific publications

This requires leadership and can be accelerated through new publishing models and compliance with standards set.

2. A fundamentally new approach towards optimal reuse of research data

Data sharing and stewardship is the default approach for all publicly funded research. This requires definitions, standards and infrastructures.

3. New assessment, reward and evaluation systems

New systems that really deal with the core of knowledge creation and account for the impact of scientific research on science and society at large, including the economy, and incentivise citizen science.

4. Alignment of policies and exchange of best practices

Practices, activities and policies should be aligned and best practices and information should be shared. It will increase clarity and comparability for all parties concerned and help to achieve joint and concerted actions. This should be accompanied by regular monitoring-based stocktaking.

Twelve action items with concrete actions to be taken

Twelve action items have been included in this Call for Action. They all contribute to



POLITICAL ISSUES (3)



**Open
data**
is about
MORE
THAN
DISCLOSURE
it must be
“**Fair**”

- Findable
- Accessible
- Interoperable
- Reusable

<http://www.dtls.nl/wp-content/uploads/2015/11/Schermafbeelding-2016-04-01-om-14.22.46-300x188.png>

REFERENCES



- ❖ Schöpfel, J., Chaudiron, S., Jacquemin, B., Prost, H., Severo, M., Thiault, F., 2014. Open access to research data in electronic theses and dissertations: An overview. *Library Hi Tech* 32 (4), 612-627.
- ❖ Prost, H., Malleret, C., Schöpfel, J., 2015. Hidden treasures. Opening data in PhD dissertations in social sciences and humanities. *Journal of Librarianship and Scholarly Communication* 3 (2), eP1230+.
- ❖ Prost, H., Schöpfel, J., 2015. *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3*. Rapport final. Université de Lille 3, Villeneuve d'Ascq.
- ❖ Schöpfel, J., Prost, H., Malleret, C., 2015. Making data in PhD dissertations reusable for research. In: *8th Conference on Grey Literature and Repositories*, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic.
- ❖ Schöpfel, J., Juznic, P., Prost, H., Malleret, C., Cesarek, A., Koler-Povh, T., 2015. Dissertations and data (keynote address). In: *GL17 International Conference on Grey Literature*, 1-2 December 2015, Amsterdam.
- ❖ Schöpfel, J., Prost, H., Rebouillat, V., 2016. Research data in current research information systems. In: *CRIS 2016*, St Andrews, 8-11 June 2016.
- ❖ Schöpfel, J., Kergosien, E., Chaudiron, S., Jacquemin, B., 2016. Dissertations as data. In: *ETD2016*, Lille 11-13 July 2016.