Univerza v Ljubljani







Joachim Schöpfel Primož Južnič Hélène Prost Cécile Malleret Anna Češarek Teja Koler-Povh

Data and dissertations

Context

A framework for a collaborative data infrastructure
 Additional funding
 Measuring and rewarding data value
 Training experts and broaden public understanding
 Incentives for green technologies
 High-level group for future planning

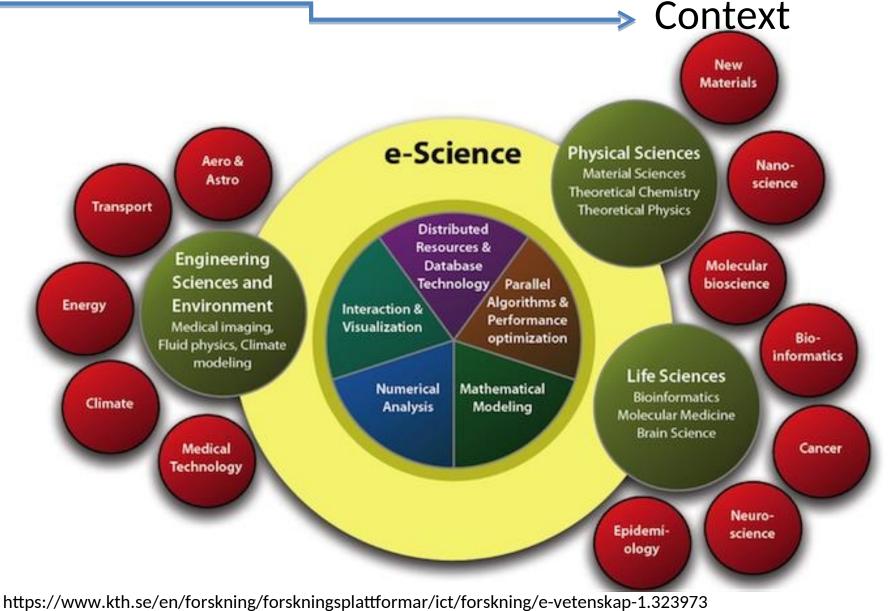
Riding the wave

How Europe can gain from the rising tide of scientific data

Final report of the High Level Expert Group on Scientific Data A submission to the European Commission

October 2010

E-science



GL17 - Dec 1-2, 2015

eScience

- Data-driven science
- Data-intensive scientific discovery

 Data deluge
- 4th paradigm (Hey et al. 2009)
 - Distributed computer and knowledge systems
 - Information and communication technologies
 - Large-scale and collaborative sciences and engineering
 - Mathematical modelling, numerical analysis, visualization techniques, data mining...
 - Integration of theories, simulations and experiments
- Scientific information has become a continuum between publication and data

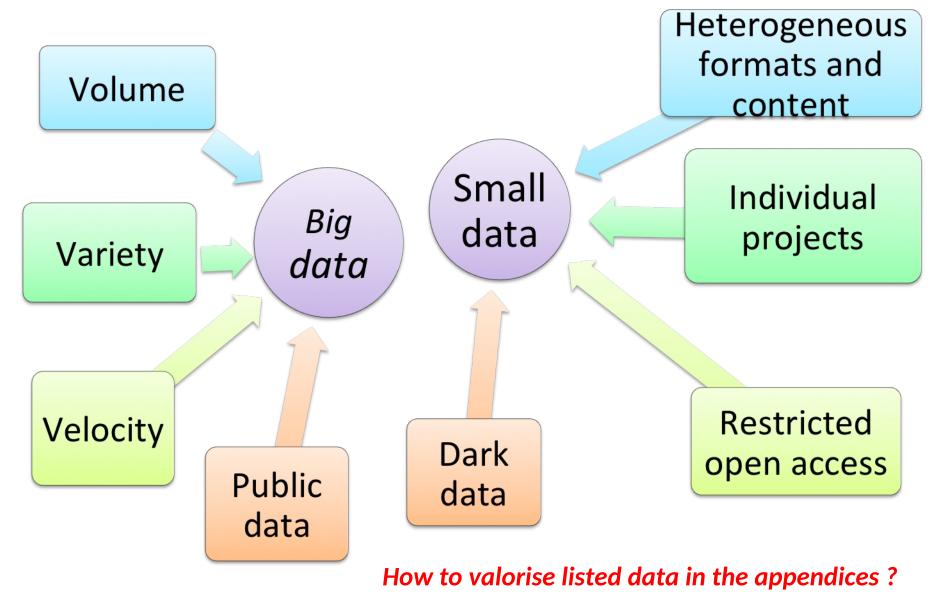
Research data

- « 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 »

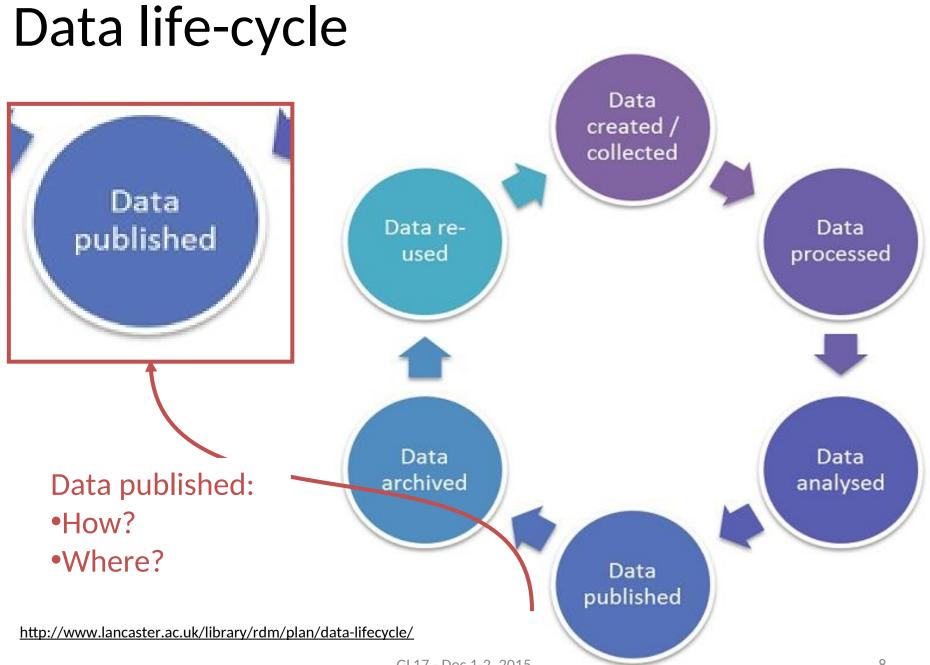
University of Edinburgh (cited by Burnham 2013)

- Large variety of formats, sources and types
- Data as material (input) v. Data as results (output)

Big data vs small data



GL17 - Dec 1-2, 2015



Data publication

Enhanced data explanation in supplementary files Publications with data

Processed Data and data representations Data published in a research article

Data referenced in research articles and available in repositories

Data in drawers

harddrives

and on researcher

Data publications, describing available datasets

Data collections and structured databases

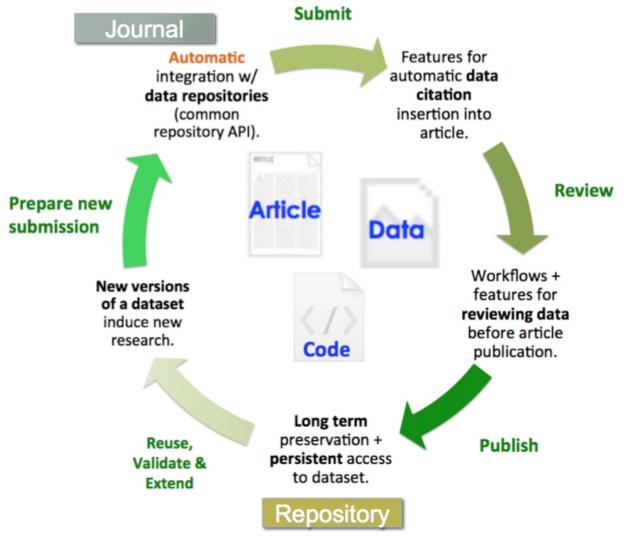
Raw data and data sets

https://www.elsevier.com/connect/can-data-be-peer-reviewed

Publication and data

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

Integration of DataVerse and OJS – An example of the gateway function



GL17 - Dec 1-2, 2015

http://journal.code4lib.org/articles/10989

Data and dissertation

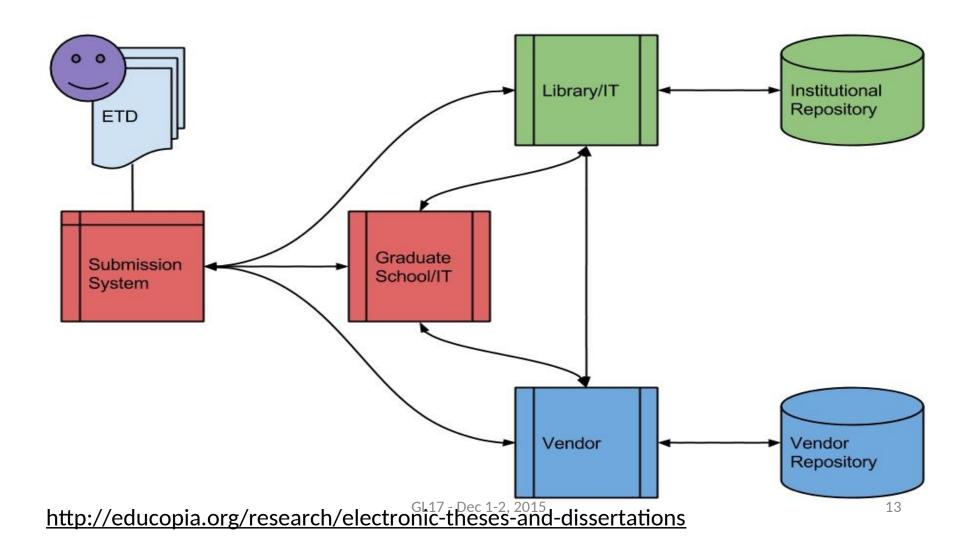
The case of ETDs

The challenge of ETDs Life-cycle management

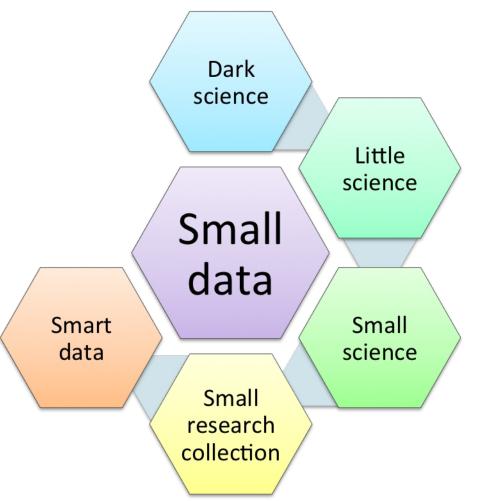
Create

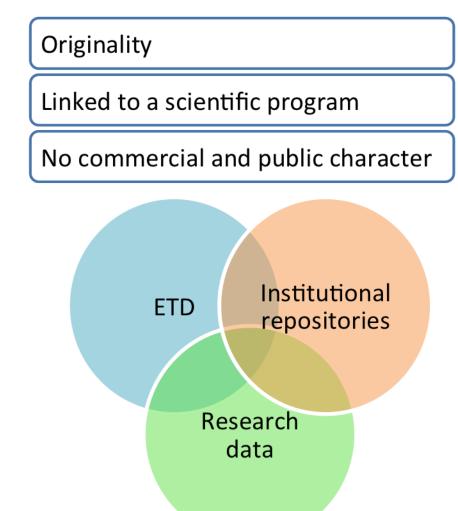
Appraise/Manage

Store

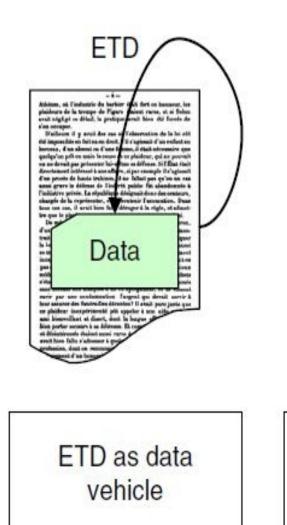


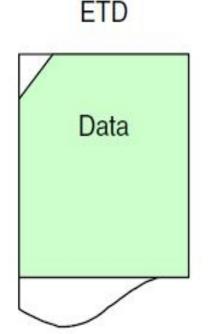
The small data of ETDs





ETDs and data





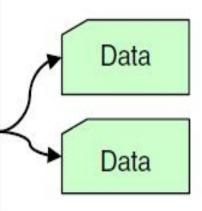


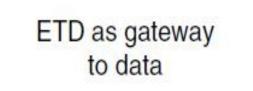


Abhom, où l'industrie de larbier flait first en inseaver, ins phildeurs de la temps de Figure shaket raves, et al Solas avait sigligt es délait, la protegue sonit liese dé la des de sin compare.

Building U p weit des cas it bisservation de la bis obt obt importable enfants aus deut. VI i s'aginant d'un enfant en horman, d'au alemat na l'aux finanz, il signi alemanites que quelqu'au partie partie aux in la source de un pinteur, qui au portant des entennes que présenve la site dans en defines. Si fillant dans des entennes que présenve la site dans en defines. Si fillant dans des entennes que présenve la site dans en defines. Si fillant dans des entennes que présenve la site dans en defines en que que a sours graves la defines du l'induite partie des présients. Trainfaire parties de l'autorité pablie dei alemanisme à l'autorité parties de l'autorité pablie de alemanisme a l'autorité parties de l'autorité pablie de alemanisme a l'autorité parties. La république désignant dons dus conteners changés de la coprésente, et de contener l'autorités parties dans leur non nas, il avrit lieux bités désagent la règit, atalantties que la piloblem es parcett la très republique que admin

ter gar la plather en parenti l'ant reprintenter par auteri. De mine, e' argemait d'un informer, d'un reporre, d'an maleite en d'an milité, deui la langes charde au mantuit relatit à la pareit, il derrect hon difficit a d'applique la d'att d'un mannet particit la tempér de super, et aust un applie d'anne que de heu produien malaideux chechanoni mangale d'anges la promise nor de son difficit d'applique par à silieurs son il plateire arisate que de parters vices moltas linearies sons il plateire arisate que des autéries auteries d'an aplicateur arisate que des mitte contains autéries autéries de la partie arise autéries autéries auteries d'an aplicateur arisate que des autéries auteries d'an aplicateur arisate que des sons défans ent athépes d'a un que plate, et a visant revier par aus malements l'auteries. La promise que en plateur insupéries de la parter a une atteire sin bismelliste ent la barrarde action la barrar auter soit blate information action la barrar auteries addition responses auxies la bismelliste action la barraries addition responses auxies la bismelliste action la barrar auteries addition responses auxies la bismelliste action la barraries addition responses auxies la bismelliste action la barraries addition responses auteries de la apprese addition responses auxies la bismelliste action la barraries addition responses auteries de la apprese addition de la mannese addition de la apprese and histories de la apprese addition de la





The potential of ETDs

- Contain the results of at least three years of scientific work
- Variety and richness of appendices
- Availability in open access
- Contribute to eScience

Data and dissertations

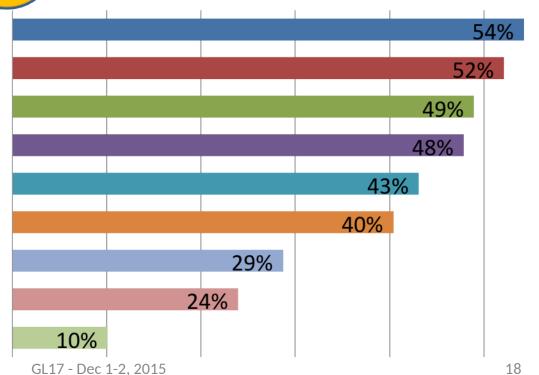
Empirical results

Data management : practices & needs Survey at Lille 3

83% on private computer 49% on professional computer

97%

declare themselves responsible for data backup

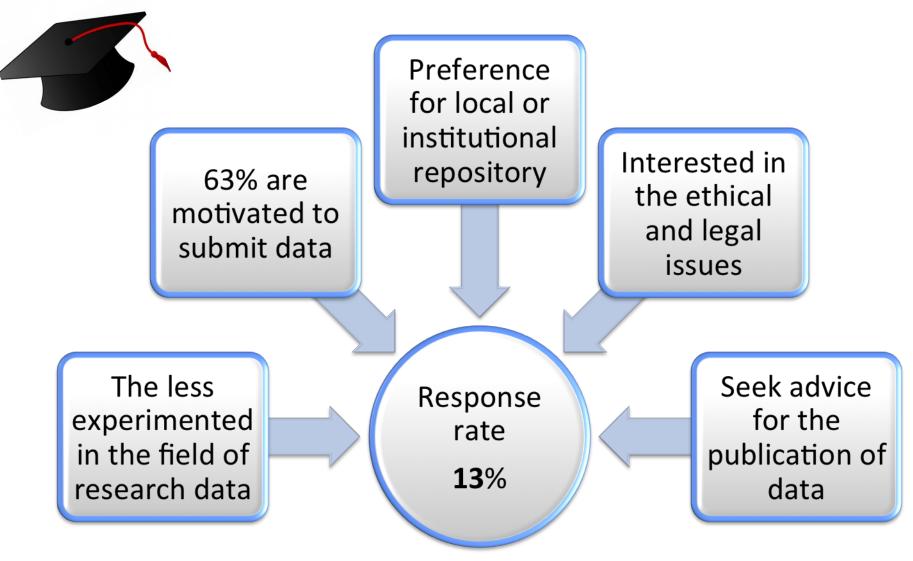


An area of storage and archiving General advice **Technical advice** Legal advice Advice- citations publications Technical assistance Assistance in preparing DMP Advice - ethics Others - no need

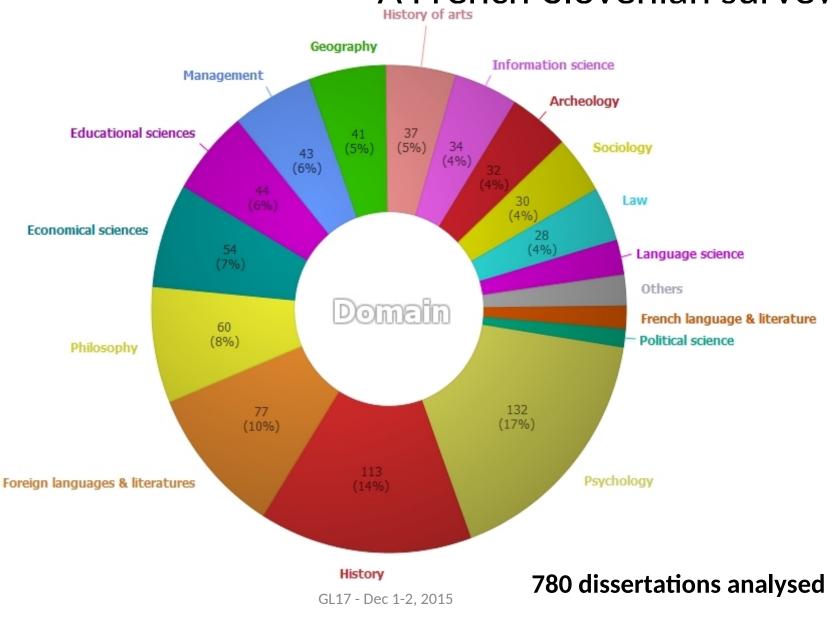
9/10

store their data on local

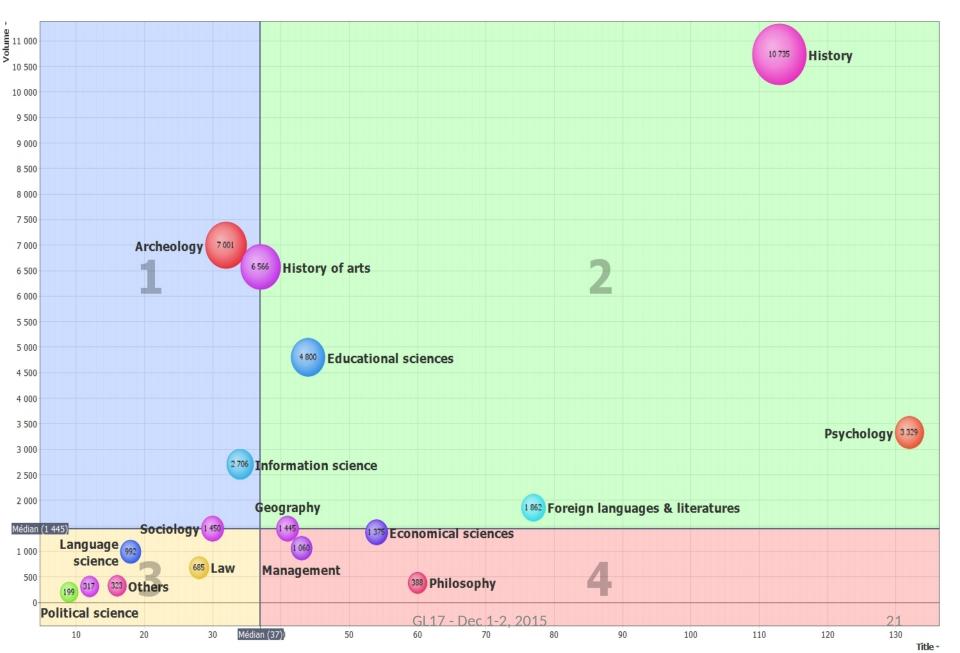
Data management : practices & needs The PhD students



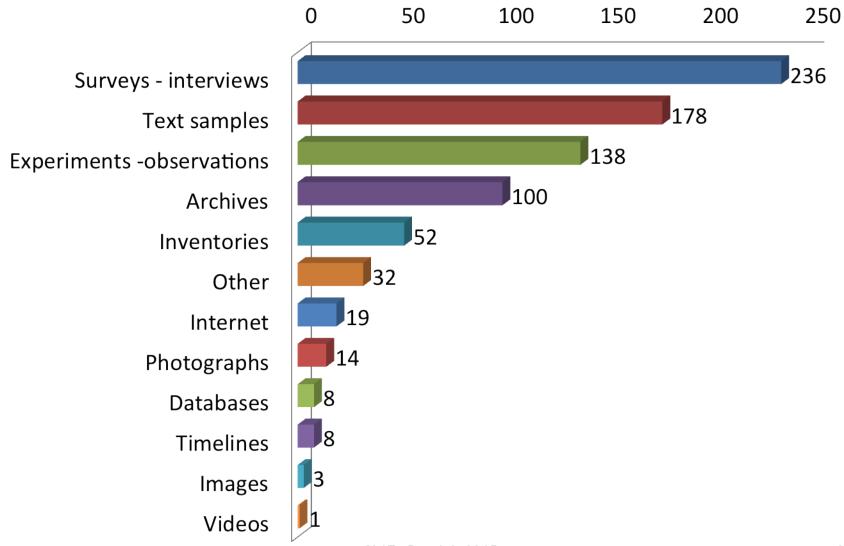
Research data in dissertations > A French-Slovenian survey



The size of appendices



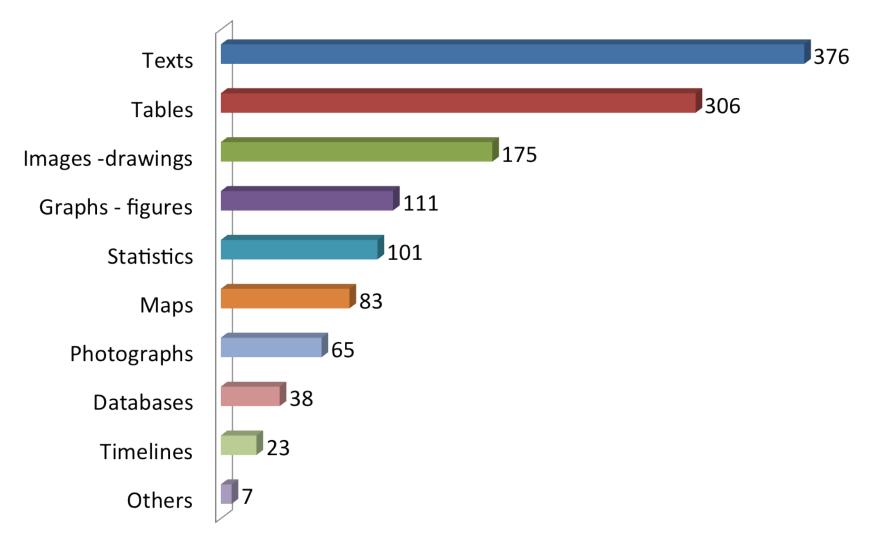
Sources of appendices



Sources of appendices

	Domains												
Signal		Experiments Surveys - Text									т. н. Г		
°.	Archives	Databases	- observati	Images	Internet	Inventories	other	Photographs		samples	Timelines	Vidéos	Total
Archeology	1		14			25		5		3			30
Economical sciences		1	27		3	1			14	18			43
Educational sciences			6		3			1	33	13	2		38
Foreign languages & literatures	4		5		1		11		14	35	1		46
French language & literature									2	3	1		6
Geography	3	2	27		1		5		15	8		1	33
History	76		4	1		7		2	9	21	2		88
History of arts	10		4			17	7	2	3	4	2		28
Information science	1	1	1		6		6		17	10			28
Language science			1						3	5			7
Law					2				2	5			7
Management	1	3	7		1	1			21	14			30
Others		1	4						1				6
Philosophy	1		2					1	1	9			11
Political science									6	3			6
Psychology			30				3	2	71	12			91
Sociology	3		6	2	2	1		1	24	15			28
Total	100	8	138	3	GL17 -19	ec 1-2, 5 20	15 32	14	236	178	8	2	3 526
													X: Sources -

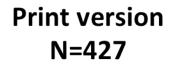
Typologies of appendices



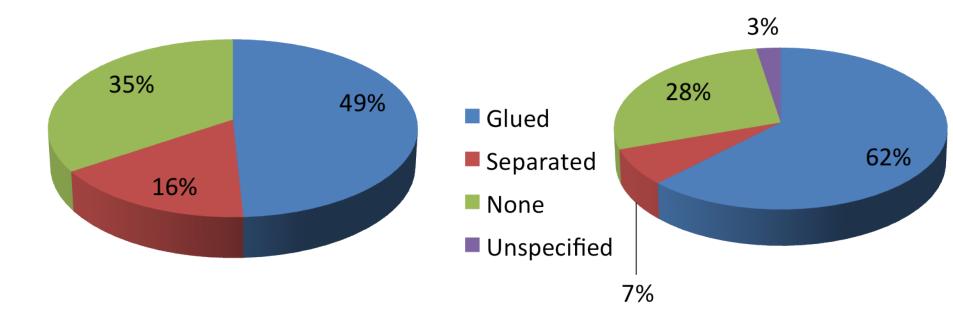
Typologies of appendices

•	Domains										
ur Mrcheology	Databases	Graphs - figures	drawings	Maps	Others	Photographs	Statistics	Tables	Texts	Timelines	Total
ă ₩ 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
Total	38	111	175	GL17 83	- Dec 1-2, 2 7	:015 65	101	306	376	23	25 526

Link between text and data appendices



Electronic version N=353



Data and dissertations

observations

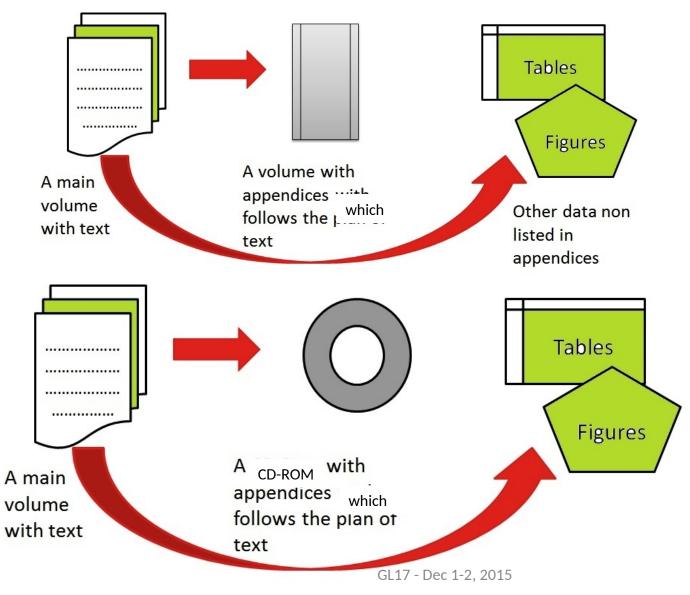
Text and data

Structure & presentation

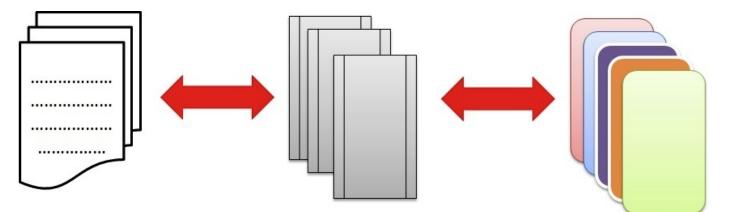


It's like bicycles !

Text and data Text and appendices separated (1)



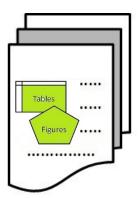
Text and data Text and appendices separated (2)



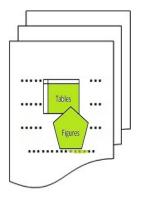
A main volume with text and classification plan A volume of appendices including catalog with 179 notices

A volume of illustrations with 1581 images

Text and data Text and data attached



Appendices are inserted at end of each chapter Other data are included in text and classified



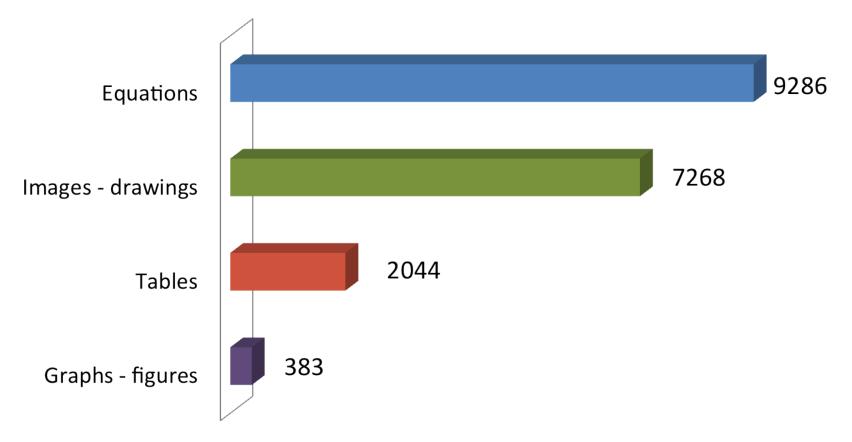
No appendices

All data are included in text with classification or not

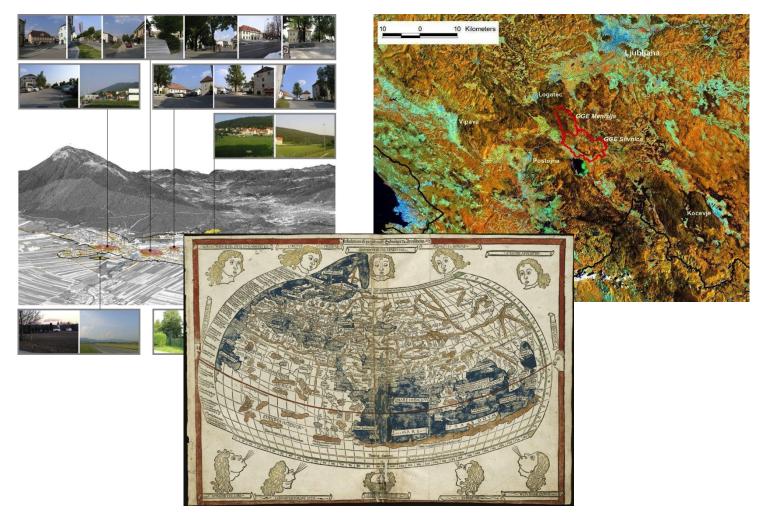
ETDs in engineering science A case study from IR DRUGG

- All dissertations (100 %) are archived in the DRUGG.
- In period 2008-2014 = 86 dissertations.
- 18,981 appendices integrated in all 86 dissertations
- also 237 attached appendices in 28 dissertations
- All content described by metadata
- Most data in format pdf

DRUGG - typologies of appendices



Photo? Map? Figure!



GL17 - Dec 1-2, 2015

The high frequenced appendices in engineering

1. Equitations (total = 9,286).

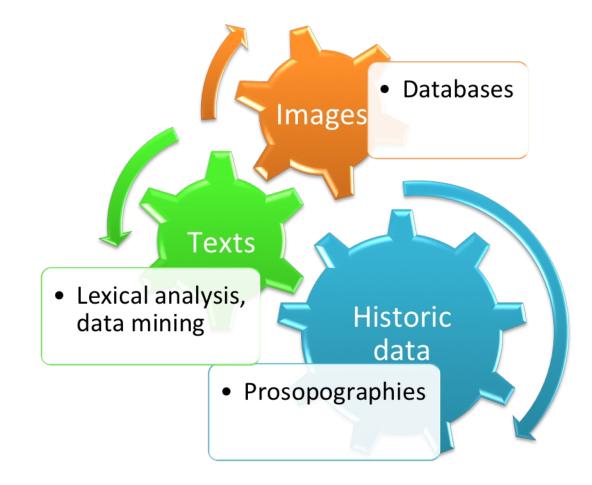
They are a part of 65 dissertations (from 86).

2. Figures (total = 7,591).

They are presented in all of 86 dissertations.

- **Specialty:** Maps and photos (they are many) are treated as figures!
- **Conclusion:** The appendices are important source of information in engineering.

Potential use & valorisation of data



Barriers to open data

Incomplete, inadequate or missing description

Missing organisation

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

Inadequate

format

Data and dissertations

Recommendations

Recommandations

Clear separation of text and data. Digital research data must be submitted in different and separate files

Separation

The data must be described in a standard language and format, with sufficient detail for retrieval and data mining

Metadata

Structuration of the research data, with a detailed and organized tagging (markup) of the datasets

Structuration

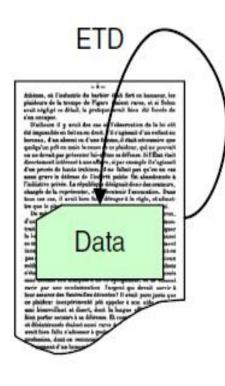
Deposit in open and original format to facilitate long-term preservation and reuse

Format

Clearing of privacy and copyright issues

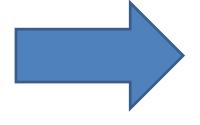
Copyright

ETDs and data



ETD as data

vehicle

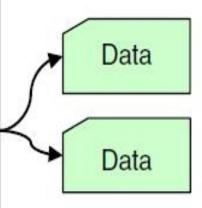


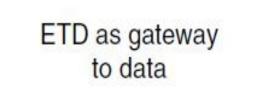


Abbiens, où l'industrie de barbier Halt fort en inseneur, tes phildeurs de la treupe de l'igner Halten raves, et si Soles avait sigligit es délait, la protegue avait bien dé farede de sin compaye.

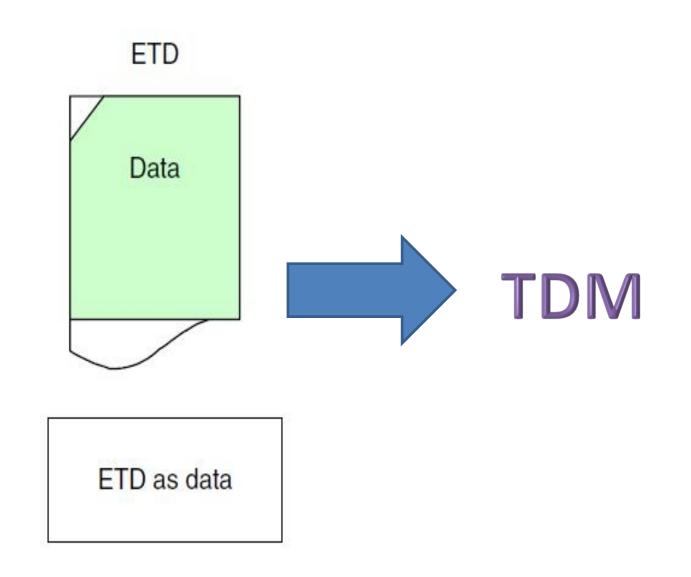
Buildone U p exit des cas à l'intervention de la loi oté del impossible enfante en de del 17 è optimul d'un estada terportipiere petre muite le conce fonces, el cuit estematere que quelqu'en petre en muite le conce faitere, pier autopreset en anciente que présente histoire en définer. Sell'Etat tait deventement intérent le conce ditre, que ensangle Eriginant d'un present de laute techere. Au la faitet par qu'un en conman greve la définer de l'actérit patie fils alamétement à l'initiative présente, et de entaire l'assurations à l'initiative présente, et de entaire l'assurations. Baue teur ens cas, l'avait him fui desegue la traje, stadance les que te juicitere querellaire la constante par entre. De miner, et a supersit l'ann interprésente par entre.

Do minimo, v 2 singlement d'un influence d'un influence d'an marichet un d'an milital, d'an is la harpe des distortier en montrait relacità à la parezie, il derenanti hom difficien d'applique, la da. Cola tati mananza pardera la bioque-des pages, et anni en negaci d'ame que de harp provisione marichette deschamant insequed des que de la bioque contratte que de la partera sizes autoritates autoritate que de la partera sizes militates lisanchies sons la barrate, que dense mainte combain citatates que autoritate que de la partera sizes militates lisanchies sons la barrate, que dense mainte combain citatates con maripeten d'en vil apropheto, et a vanante estate sons consistentation. l'angeque qui derent aucoré de la parter sonsarré au diference. En que estatismente estates estates autoritates de la parter ais biomeditates estates de la parte autoritate diference autoritate estates de la parter estate destatemente de la segue parter sonsarré ta su difference. En que estates estates estates estates autoritates estates estates de la calemente a parter regularisme, dust en encourses





ETDs and data



Helping PhD students to manage data Services

- Seminars, conferences, training
- Online resources (guidelines, FAQ...)
- Alert service

Education

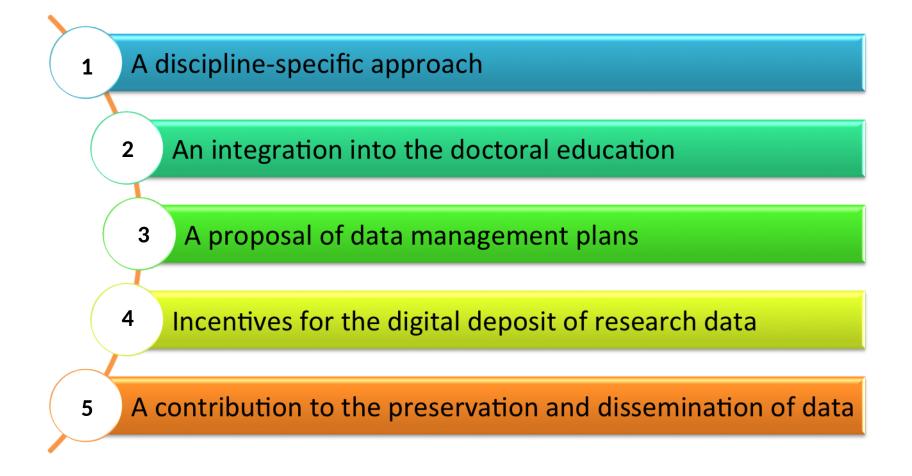
- Legal and technical help (data management plan)
- Technical assistance for deposit
- Liaison with laboratories

Advice,

assistance

- Mediation for deposit
- Partnership with networks and repositories
- Development of tools on the campus

Infrastructures



Data and dissertations

Thank you !

Contact joachim.schopfel@univ-lille3.fr primoz.juznic@ff.uni-lj.si

References http://www.citeulike.org/user/Schopfel/tag/gl17