Data deposit in a CKAN repository: a Dublin Core-based simplified workflow

Yulia Karimova, **João Aguiar Castro**, Cristina Ribeiro

INESC TEC, Faculdade de Engenharia da Universidade do Porto

15th Italian Research Conference on Digital Libraries, Pisa, 1 February, 2019

INESCTEC
TECHNOLOGY & SCIENCE

Institute for Systems and Computer Engineering, Technology and Science

Research / Clusters

- Computer Science
- Industry and Innovation
- Networked Intelligent Systems
- Power and Energy

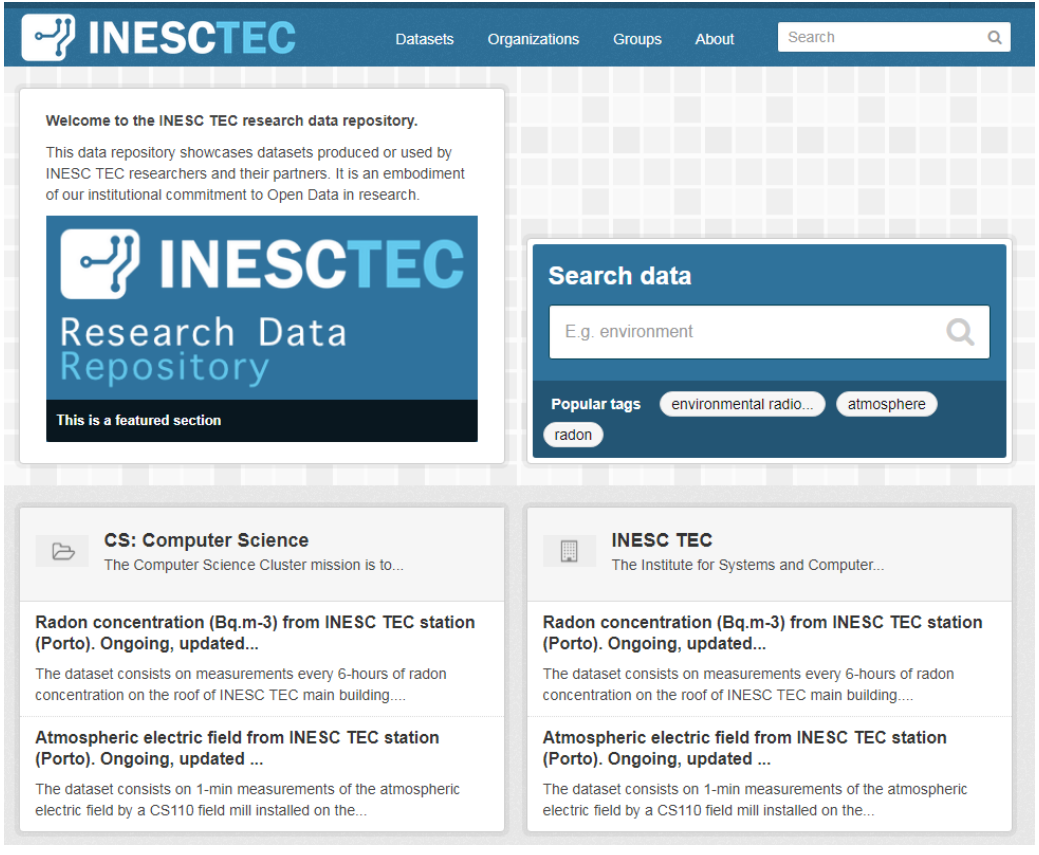
Data repository at INESC TEC - Motivation

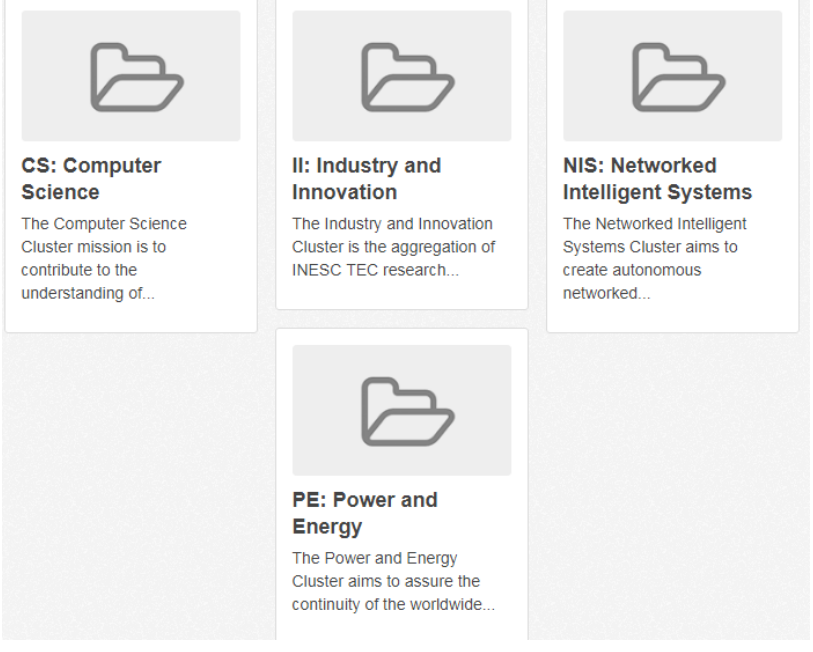
Engage researchers in RDM through data description, using Dublin Core

1) Practical when assessment of requirements is not possible;
2) Easier for researchers to grasp concepts behind Dublin Core descriptors

Many datasets from closed projects will not reach publication stage without an agile deposit process

INESC TEC data repository





INESC TEC data repository

Albergaria Fire Fighters ECG 2010

Monitoring of Albergaria Firefighters ECG during emergency events during the summer of 2010 and correspondent event labels.

Data and Resources

SESSION-2010-06-07-01-OUTSIDE.zip
Approximately 59 hours of ECG missions recordings of 5 Firefighters in .bdd...

Explore

AlbergariaFirefighters

Additional Info

Field	Value
Author	Joana Paiva, João Paulo Cunha
Last Updated	4 de Dezembro de 2018, 18:22 (UTC+00:00)
Created	19 de Janeiro de 2017, 13:23 (UTC+00:00)
DOI	https://doi.org/10.25747/NPNT-P555
dc.Contributor	Vital Responder 1.0 team, VR2Market team
dc.Coverage.Spatial	Albergaria
dc.Coverage.Temporal	February-July 2010

dc.Date	Summer, 2010
dc.Format	*.bdd, *.xls
dc.Format.Extent	165MB
dc.Identifier	2010_AlbergariaFFs_ECG
dc.Publisher	INESC TEC
dc.Relation	Brás, S., Fernandes, J. M., & Cunha, J. P. (2013, June). ECG delineation and morphological analysis for firefighters tasks differentiation. In Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on (pp. 516-517). IEEE. ISO 690 DOI: 10.1109/CBMS.2013.6627856; Pallauf, J., Gomes, P., Brás, S., Cunha, J. P. S., & Coimbra, M. (2011, August). Associating ECG features with firefighter's activities. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 6009-6012). IEEE DOI: 10.1109/iembs.2011.6091485
dc.Type	ECG data, Events labels

Descriptors in the repository

Dataset attributes	Corresponding descriptor and vocabulary
Availability	Visibility (CKAN), DOI
Bibliometric data	-
Coverage	Coverage.Temporal (Dublin Core), Coverage.Spatial (Dublin Core)
Date	Date (Dublin Core)
Format	Format (CKAN), Format (Dublin Core), Format.Extent (Dublin Core)
License	License (CKAN); License (Dublin Core)
Minimal description	Title (CKAN), Name (CKAN), Author (CKAN), Author email (CKAN), Description (CKAN), Maintainer (CKAN), Maintainer email (CKAN), Type (Dublin Core), Language (Dublin Core), Publisher (Dublin Core), Contributor (Dublin Core)
Paper reference	Relation (Dublin Core)
Project	Organization (CKAN), Group (CKAN)
Provenance	Source (CKAN), Version (CKAN)
Subjects	Tags (CKAN)

Deposited datasets

Domain	Datasets deposited in 2017				2018	
	1 trimester	2 trimester	3 trimester	4 trimester	1 trimester	2 trimester
Biomedical engineering	•					
Environmental radioactivity	•	•	•	•	•	•
Biomedicine		•			•	
Robotics					•	
Information science					•	
Natural language processing		•	•	◇		
Music streaming	•					
Information retrieval	•		◇			

• – *public datasets* ◇ –*private datasets*

Description results

Descriptor	Descriptors used by researchers	Descriptors after curator feedback
Title	21	21
Author	21	21
Author email	21	21
Description	20	21
Format	20	21
Tags	19	21
License	16	21
Coverage (Temporal)	16	17
Type	16	20
Date	15	21
Coverage (Spatial)	14	15
Language	14	16
Publisher	12	21
Relation	12	15
Source	10	10
Contributor	6	6
Format.Extent (File size)	5	21

General conclusions

- Metadata and descriptor concepts are not easy to understand;
- ***Contributor*** only for external parties; ***Author*** for all group members;
- More detailed metadata after data curation feedback; ***Title*** often revised.

General conclusions

- Researchers do not want to spend much time in description;
- Inconsistent metadata will improve with controlled vocabularies (*e.g.* ISO 639-2);
- Curiosity about data citation and sense of awareness regarding data preservation.

Researchers feedback

"*We have **used the dataset link in our papers***"

"*I think our data was not reused yet (or, if they were, we are not informed; I**t would be useful to have some idea of how many times the data was downloaded***";

"*At first I found it hard, but as we go it becomes **much easier and simpler***"

*"The process itself is **a bit time consuming**. The choice of descriptors is not something for which we are oriented and **the support of the curator is fundamental here**"*

Conclusions

- A first step to involve researchers in data description;
- Different levels of RDM awareness require different approaches and flexibility;
- Experienced researchers showed greater interest - collaboration in other RDM activities, such as the development of DMPs
  (although expectations are not quite fulfilled);
- Metadata is still lackluster and there is a need to extend the metadata form.