Data Management Plan

FIRM-CENTRIC TECHNOLOGY ROADMAPS (FTRMS) AND TECHNOLOGY VALUATION – Bruno Cassiman (G010324N)

1. General Project Information	
Name Grant Holder & ORCID	**Bruno Cassiman (0000-0001-9602-6755)**
Project number & title	G010324N FIRM-CENTRIC TECHNOLOGY ROADMAPS (FTRMS) AND TECHNOLOGY VALUATION
Affiliation(s)	■ KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: ROR identifier KU Leuven: 05f950310
Please provide a short project description	The project proposes to develop Firm-centric Technology Roadmaps (FTRMs) and to improve technology valuation models based on them. Anticipating technology evolutions has become a critical business skill. Evaluating a potential technology opportunity in relation to the organization's own technological capabilities is critical for success. We address three gaps in the literature. First, we develop a methodology for quantifying Technology Roadmaps (TRMs) and white spaces therein based on patent and publication information. Second, we link these TRMs to critical firm-specific scientific and technological capabilities. Third, we develop a method to estimate the economic value of technologies based on these FTRMs. Altogether, this will allow answering important economic and business strategy questions related to technology development and entry-timing decisions of the organization. To this end, we leverage recent Natural Language Processing and Machine Learning methodologies based on patent and publication text.

2. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				ONLY FOR DIGITAL DATA	*ONLY FOR DIGITAL DATA*	*ONLY FOR DIGITAL DATA*	*ONLY FOR PHYSICAL DATA*
Dataset Name	Description	New or Reused	Digital or Physical	Digital Data Type	Digital Data Format	Digital Data Volume (MB, GB, TB)	Physical Volume
Patentsview data	Descriptive and full text data from USPTO patents.	☐ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	.csv Python file	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	/
PATSTAT	More comprehensive descriptives and full text data on patents (beyond USPTO)	Reuse existing data	Digital	Numerical and textual	.csv Python file	<1 TB	/
KPSS data	Results of commercial value of patents from the findings of Kogan et al. (2017)	Reuse existing data	Digital	Numerical	.dta Python file	<100 GB	/
DISCERN data	database of publicly listed	Reuse existing data	Digital	Numerical	.dta	<100 GB	/

	U.S. headquartered firms matched to assignees of patents from the United States Patent and Trademark Office (USPTO) and scientific publicationsfrom the Web of Science for the period 1980-2015.						
Orbis	We use financial and firm information from the Orbis historical database. (which covers a large chunk of non listed firms)	Reuse existing data	Digital	Numerical	.csv Python file	<100 GB	/
Compustat	We use financial and firm information from compustat, provided by WRDS	Reuse existing data	Digital	Numerical	.csv Python file	<1 GB	/

AHG data	We use data from Arts, Hou and Gomez (2020) on novelty of patents	Reuse existing data	Digital	Numerical and Textual	.cvs	<100 GB	/
OpenAlex	We use data on Scientific publications from OpenAlex. OpenAlex is a bibliographic catalogue of scientific papers, authors and institutions accessible in open access mode	Reuse of existing data	Digital	Numerical and Textual	JSON (database)	<100 GB	/
Revenue and Patents data IMEC	We use data on contract revenue related to patents of IMEC	Reuse of existing data	Digital	Numerical	.cvs Python file	<100 GB	/

GUIDANCE:
The data description forms the basis of your entire DMP, so make sure it is detailed and complete. It includes digital and physical data and encompasses the whole spectrum ranging from raw data to processed and analysed data including analysis scripts and code. Physical data are all materials that need proper management because they are valuable, difficult to replace and/or ethical issues are associated. Materials that are not considered data in an RDM context include your own manuscripts, theses and presentations; documentation is an integral part of your datasets and should described under documentation/metadata.
RDM Guidance on data

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type.	Patentsview (csv): https://patentsview.org/download/data-download-tables PATSTAT: https://inspire.wipo.int/patstat-online KPSS (dta): https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Replication-Kit ORBIS: https://bib.kuleuven.be/ebib/collectie/data/databanken/orbis_global DISCERN: https://zenodo.org/records/4320782 AHG: https://zenodo.org/records/3515985 OpenAlex: https://openalex.org/
Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.	☐ Yes, human subject data; provide SMEC or EC approval number: ☐ Yes, animal data; provide ECD reference number: ☐ Yes, dual use; provide approval number: ☒ No Additional information: The data we use is either publicly available or licensable for any party (give a fee applies).
Will you process personal data[1]? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).	☐ Yes (provide PRET G-number or EC S-number below) ☒ No Additional information: We will strictly only use firm and patent information. Individual inventors will not be analyzed in this project.
Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.	☒ Yes ☐ No If yes, please comment: We only use (semi)publicly available data which depending on the results could be used in the deployment of a tool to predict the value for patents and related whitespaces in technology and technology portfolios. To this end we will use the public (and licensed) data to train our models but relate the predictions to each firms proprietary data in the final version of the spin-off. (Which is beyond the scope of this project)

[1] See Glossary Flemish Standard Data Management Plan

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements, research collaboration agreements)? If so, please explain to what data they relate and what restrictions are in place.	☒ Yes ☐ No If yes, please explain: The Orbis and WRDS data has to be licensed by the KUL, currently they have a good relationships and they are licenses which the KUL is eager to extend. However, if at any point Orbis restricts access the KUL will have to destroy their copies and SQL application on Orbis. The access towards WRDS could also be restricted. However, in both those cases I would have to secure my own access, potentially through the use of a Bench Fee.
Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain to what data they relate and which restrictions will be asserted.	☒ Yes ☐ No If yes, please explain: Since we are all in order with our licensed data (Orbis/WRDS) being licensed correctly and all other data being publicly available there is no issue with regards to the usage of this data for this research project. However, in the future possible commercial exploitation some new licenses will probably have to be negotiated in order to comply with the regulations from the data providers. We will also have access to proprietary data from IMEC.

3. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep **data understandable and usable**, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded). *RDM guidance on documentation and metadata*.	I will provide a file that shows which data was collected from where and how it relates to each project. Furthermore, in each of the coding files the sources are mentioned and I will provide comments in each step in order to further enhance the reusability of the code as well as help with versioning. When publishing the code there will be README.txt files provided for each of the important steps to enhance replication. Replication will be made as easy as reading in the files sequentially to get the results (given that the data paths are adjusted of course).
Will a metadata standard be used to make it easier to **find and reuse the data**? If so, please specify which metadata standard will be used. If not, please specify which metadata will be created to make the data easier to find and reuse. *REPOSITORIES COULD ASK TO DELIVER METADATA IN A CERTAIN FORMAT, WITH SPECIFIED ONTOLOGIES AND VOCABULARIES, I.E. STANDARD LISTS WITH UNIQUE IDENTIFIERS.*	☒ Yes ☐ No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: A README.txt file will provide the necessary context for the data as well as specify the 2 unique identifying variables which we will use. Firstly, we have Patstat's patent_id (As well as patentsview patent_id). Seconldy firms are identified with the unique GVKEY provided by WRDS computstat. These unique identifiers should be enough to link all of our data together. Further data clarifications on potential variables we will develop will also be included in the README.txt file.

4. Data Storage & Back-up during the Research Project

Where will the data be stored? *Consult the interactive KU Leuven storage guide to find the most suitable storage solution for your data.*	☐ Shared network drive (J-drive) ☐ Personal network drive (I-drive) ☒ OneDrive (KU Leuven) ☐ Sharepoint online ☐ Sharepoint on-premis ☒ Large Volume Storage ☐ Digital Vault ☒ Other: MSI server (Patstat data)
How will the data be backed up? *WHAT STORAGE AND BACKUP PROCEDURES WILL BE IN PLACE TO PREVENT DATA LOSS?*	☒ Standard back-up provided by KU Leuven ICTS for my storage solution ☒ Personal back-ups I make (the source data will be backed up on the KU Leuven onedrive but results and preprocessed data is also backed up on an external SSD drive) ☐ Other (specify)
Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.	☒ Yes ☐ No If no, please specify:

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *CLEARLY DESCRIBE THE MEASURES (IN TERMS OF PHYSICAL SECURITY, NETWORK SECURITY, AND SECURITY OF COMPUTER SYSTEMS AND FILES) THAT WILL BE TAKEN TO ENSURE THAT STORED AND TRANSFERRED DATA ARE SAFE.* *Guidance on security for research data*	The data used will be kept securely on the MSI server as well as personal network drives from the KUL. These are regularly updated and maintained by the FEB ICT personnel as such they are some of the default options at the KUL. When collaborating with others I will use the Shared network drive :J, these can only be accessed via VPN connection to the KU Leuven network and follows the same maintenance and update scrutiny from the ICT personnel.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	If any further data storage is required we will tap into the bench fee provided by FWO for the yearly workings of this project. As such there will always be budget in the case that storage costs would suddenly increase.

5. Data Preservation after the end of the Research Project	
Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...). *Guidance on data preservation*	☒ All data will be preserved for 10 years according to KU Leuven RDM policy ☐ All data will be preserved for 25 years according to CTC recommendations for clinical trials with medicinal products for human use and for clinical experiments on humans ☐ Certain data cannot be kept for 10 years (explain)

Where will these data be archived (stored and curated for the long-term)? *Dedicated data repositories are often the best place to preserve your data. Data not suitable for preservation in a repository can be stored using a KU Leuven storage solution, consult the interactive KU Leuven storage guide.*	☒ KU Leuven RDR ☐ Large Volume Storage (longterm for large volumes) ☒ Shared network drive (J-drive) ☐ Other (specifiy):
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	The expected costs would the standard costs related to storing projects on KU Leuven RDR, this would mean that no extra costs would be imposed on this project.

6. Data Sharing and Reuse	
Will the data (or part of the data) be made available for reuse after/during the project? Please explain per dataset or data type which data will be made available. *NOTE THAT 'AVAILABLE' DOES NOT NECESSARILY MEAN THAT THE DATA SET BECOMES OPENLY AVAILABLE, CONDITIONS FOR ACCESS AND USE MAY APPLY. AVAILABILITY IN THIS QUESTION THUS ENTAILS BOTH OPEN & RESTRICTED ACCESS. FOR MORE INFORMATION: HTTPS://WIKI.SURFNET.NL/DISPLAY/STANDARDS/INFO-EU-REPO/#INFOEUREPO-ACCESSRIGHTS*	☐ Yes, as open data ☐ Yes, as embargoed data (temporary restriction) ☒ Yes, as restricted data (upon approval, or institutional access only) ☐ No (closed access) ☐ Other, please specify:
If access is restricted, please specify who will be able to access the data and under what conditions.	

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain per dataset or data type where appropriate.	☐ Yes, privacy aspects ☒ Yes, intellectual property rights ☐ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ☐ No If yes, please specify: As part of the licensing agreement we are not allowed to share the Orbis data to third parties, the results of our analysis can be published freely. The data on contract revenue of IMEC is confidential and shared with the project, but not available for sharing.
Where will the data be made available? If already known, please provide a repository per dataset or data type.	☒ KU Leuven RDR ☐ Other data repository (specify) ☐ Other (specify)
When will the data be made available?	☒ Upon publication of research results ☐ Specific date (specify) ☐ Other (specify)

Which data usage licenses are you going to provide? If none, please explain why. *A DATA USAGE LICENSE INDICATES WHETHER THE DATA CAN BE REUSED OR NOT AND UNDER WHAT CONDITIONS. IF NO LICENCE IS GRANTED, THE DATA ARE IN A GREY ZONE AND CANNOT BE LEGALLY REUSED. DO NOTE THAT YOU MAY ONLY RELEASE DATA UNDER A LICENCE CHOSEN BY YOURSELF IF IT DOES NOT ALREADY FALL UNDER ANOTHER LICENCE THAT MIGHT PROHIBIT THAT.* *Check the RDR guidance on licences for data and software sources code or consult the License selector tool to help you choose.*	☐ CC-BY 4.0 (data) ☐ Data Transfer Agreement (restricted data) ☐ MIT licence (code) ☐ GNU GPL-3.0 (code) ☐ Other (specify)
Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *INDICATE WHETHER YOU INTEND TO ADD A PERSISTENT AND UNIQUE IDENTIFIER IN ORDER TO IDENTIFY AND RETRIEVE THE DATA.*	☒ Yes, a PID will be added upon deposit in a data repository ☐ My dataset already has a PID ☐ No
What are the expected costs for data sharing? How will these costs be covered?	

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	**Bruno Cassiman**
Who will manage data storage and backup during the research project?	**Bruno Cassiman**
Who will manage data preservation and sharing?	**Bruno Cassiman**
Who will update and implement this DMP?	**Bruno Cassiman**