Framework document Steering group VSNU/NFU/NWO – Elsevier

Name Pilot/Service: Research Infrastructure tracking service
Why this pilot: Research Infrastructures (RI) are an essential component of research, critical both for innovation and economic development. Tracking RI contributes to improving reproducibility of research, an essential part of Open Science. Researchers need to know what equipment was used for a given experiment, and how to access it or what alternative equipment to use to reproduce an experiment. This can be made possible when research outputs are automatically linked to the equipment used to produce that research. This data, powered by a rich Research Infrastructure ontology, would support Open Science, and reproducibility in particular, by allowing researchers to identify substitute/alternative equipment to replicate experiments, or find the most convenient partner that owns a certain instrument. What will the pilot project initially focus on: There is currently no solution available that allows stakeholders to address this issue. This pilot project therefore makes the first step. The envisioned solution will be based on a full-text mining approach (at a later stage informed by an open ontology of research infrastructures), enhanced by linking with other internal and external data sources, thus connecting research infrastructures and research outputs. Elsevier and interested Dutch institutions collaborate in this pilot project to validate the feasibility of the full-text mining approach, as well as further confirming the value/usefulness of the solution. The project will be organized as follows: • The institution will provide a list of RIs that they are interested in tracking, as well as a small sample of publications and/or researchers with relations to specific items in the RI list • Elsevier will put together a corpus of publications from the institution and Elsevier's team of Data Scientists will apply the text mining algorithm to find instances of items in the RI list • Meanwhile, Elsevier will work with stakeholders at the institutions to define a set of analytics / reports of interest to the participating institutions. • At the end of the project (the estimated duration of the project should be 5 to 7 months in total, with a certain number of iterations to refine the outputs), the institution will have the reports as well as the raw data. Feedback on how the institution uses this report would also be interesting to further refine our deliverables • The collaboration will require some time commitment from the institution (maximum 1 to 2 working days per month to provide input at the beginning of the pilot and review outputs at each iteration)

1. (a) Participating institutions Participation in the Professional Services is at each Institution's sole discretion and a pilot shall only commence if there is a minimum participation by at least three Institutions *	Evaluation YES NO	Evidence and Comments
Are at least 3 institutions involved in the pilot?	YES	TU Delft, TU Eindhoven, VU Amsterdam, Wageningen University and Research
Evidence of how and when other institutions can join	YES	Any institution can join the pilot at any point in time.
1. (b) Interoperability and vendor neutrality Elsevier shall use all reasonable efforts to ensure that the Professional Services are interoperable, both on the input (uploaded) and output side (created) *	Evaluation YES NO	Evidence and Comments RI references from will be mined from all sources indexed in Scopus, meaning over 7,000 publishers worldwide On the output side, the data will be made available in standard formats
Use of open identifier systems	YES	Publications where equipment is mentioned are referenced via standard identifiers (DOIs) as well as proprietary ones (Scopus document IDs), in order to maximize the possibility of linking the data. When building an open ontology of research infrastructures, partners and Elsevier will liaise with other bodies who are looking into open PIDs for RIs and will use open PIDs whenever these are available.
Use of standardized metadata schemas	NA	The field is very new, but we are aware of community initiatives such as https://www.rd-alliance.org/groups/persistent-identification-instruments-wg

Existence of a well-documented API and open data-dump function	NOT in this phase of the pilot	In this phase of the project Elsevier will provide data dumps. Elsevier's long-term plan includes building an API
Ability to export data in a variety of formats	YES	Elsevier will provide exports in Excel / CSV format
Ability for other commercial parties to join	YES	Equipment vendors are welcome to get involved.
2. Transparency, inclusion and collaboration The Services and resulting Deliverables are aimed to make science and research more transparent, efficient, inclusive, openly and freely accessible, and collaborative. *	Evaluation YES NO	Evidence and Comments With the objective to make the whole research process open and transparent, the RI currently is an important missing piece of the research process. This pilot would provide greater transparency regarding how RI supports the production of research.
Provenance on how and where metadata was derived	YES	Metadata are collected from publications and possibly from institutions. Metadata collected from institutions will not be stored and shared beyond the scope of the pilot project without explicit consent from the institution

Descriptions of workflows that result in indicators, metrics and/or other relevant outcomes will be open and transparent. These will demonstrate, for example calculation steps, search strings used to define entities, etc.	NA	Research Infrastructure is identified by text mining scientific articles using Machine Learning models that we plan to instruct with the support of the institutions. Once publication-RI links are identified, all metrics currently available for publications and related entities (data, grants, patents, ...) may become available for RI as well. The processing pipeline includes three steps: • Identify sentences containing mentions of RI • Within those sentences, isolate the strings that constitute a reference to a RI • Link that reference to an existing knowledge base of RI If this pilot is successful, the vision is to develop an open ontology of RIs (not in the scope of this pilot)
Description of the services used to create metadata	YES	Text mining from full text and potentially curation from project participants
Insights and lessons published with Open Access license	YES	Pending all partners' approval, a report detailing these insights and learnings will be shared openly
Will the pilot contribute to Open Science?	YES	This project will enrich research data / bibliographic information about research infrastructures. The data can be released openly, and the recipe behind the algorithm will be openly shared. In a follow up of this pilot phase partners aim to develop an open ontology of research infrastructures.

		Such an ontology would greatly support research reproducibility (for example helping researchers to identify equivalent RIs when they don't have access the ones mentioned in research publications).
Demonstration of connection to non-Elsevier products	YES	We will link equipment to non-Elsevier content: at first with Scopus content (7000 publishers), but the plan is to extend to other content types in the future (preprints, research data, patents, clinical trials,..)
3. Access to research data and metadata Elsevier will give enduring access during the Term to all (research) data, including metadata, analytics and information*	Evaluation YES NO	Evidence and Comments This is the main deliverable of the project, and it will be delivered to the institution as files
Describe the ownership / licensing of data made as part of the service	YES	Once the generated data has been shared with the participating institutions, they will have the possibility to put into their systems of record. Institutions can then decide what do with this data.
Describe how access (institutional and / or public) to the data will be set-up during the term; this section will also indicate cases where certain data is not publicly access.	YES	The data will be delivered as files. During the project, Elsevier may grant access to “admin” users of the participating institutions to tools that facilitate the task of data annotation and curation
4. Data portability Institution shall be entitled to transfer the data provided, uploaded or created to its own or to a third-party host environment *	Evaluation YES NO	Evidence Comments The data will be delivered as a set of files, with a license that grants the institution the rights to host such data where they prefer
Evidence on how data transfer is possible.	YES	Data are files
How can an institution withdraw data?	NA	Data are files

5. Intellectual property *	Evaluation YES NO	Evidence Comments
Details on IP related to data provided, created or enriched	NA	NA
6. Additional considerations	Evaluation YES NO	Evidence Comments
What processes will be put in place to evaluate the service during and at the end of pilot	YES	An iterative approach to improve the accuracy and completeness of the matching algorithm will be implemented. Performance will be measures with standard metrics such as precision, recall and F1 score. The evaluation will be done against the training data provided by the partners and Elsevier.
Terms of use of the deliverables during and after contract period	YES	The deliverables of this pilot are stated in the SoW. All data serviced during the pilot will remain with the partners when the pilot is ended and or continued into a new phase or a service. Any future product or service that will be developed based on this pilot will be made available to pilot participants for the duration of the overall agreement. This includes all means to use or consume this service, such as an online user interface or an API.
Pilot project team	YES	Names of the partners' contacts will be detailed in the respective SoWs

** For the full text, please refer to the contract.*

Approved by the VSNU/NFU/NWO-Elsevier steering group	Date: