Data trusts: A techno-legal approach for sustainable data commons

> **I. Thesis proposal of Ms. Ramya Chandrasekhar, PhD candidate in public law at Paris 2 – Panthéon Assas employed at CIS CNRS UPR 2000 (1 page max)**

There is little doubt that open data serves the public interest. Open data are datasets and creative content that can be shared and reused with little to no technical or legal restrictions. And open data has been central to open government, open science and free culture.

But open data are also facing sustainability issues. The capabilities to realise value from open data are not distributed equally, resulting in overuse by some actors, such as commercial actors acting in their private interest. This results in the "paradox of open", where open access resources enable new concentrations of power and wealth, most recently visible in the use of the digital public domain and open repositories by large platform companies to train and build proprietary AI models. Further, reusers give little to no value back either to maintenance of the open access resources nor to the wellbeing of data producers (i.e. communities). Beyond simply open access to resources, there is a need for strong governance of data sharing and reuse ecosystems as a commons, to ensure equitable use and reciprocity. Ms. Chandrasekhar will study and propose "data trusts" as a legal-institutional framework for data commons, to enable sharing and reuse of digital data while ensuring sustainability of both resources *and* communities collaborating to produce and curate open datasets.

Ms. Chandrasekhar's doctoral thesis will comprise of three parts. In Part I, she will undertake a conceptual analysis of the sustainability problems of open data *as facilitated by law*. She will focus on two types of legal instruments for open data - (i) legal regulations for the public release of certain datasets and content, such as open government data, and (ii) open licenses that apply 'legal hacks', notably the inverse logic of copyright. Using theories from law and political economy, commodification studies and critical data studies, Ms. Chandrasekhar will analyse how these legal instruments construct open data as a commodity, valourising (and naturalising) the value of open data as exchange value, to the exclusion of public value and public interest. She will then analyse how such commodification by law causes frictions for equitable use of open access resources and sustainability of data producing/stewarding communities.

In Part II of her thesis, Ms. Chandrasekhar will survey the concept of "data commons" as a governance model for sharing and reuse of data. She will specifically survey new licenses for data sharing as well as new institutional structures such as data trusts, data cooperatives and data unions to organize the governance by communities. She will also analyse the opportunities and challenges that these institutional structures present for sustainability of open data.

In Part III of her thesis. Ms. Chandrasekhar will zoom into "data trusts" as a specific legal and institutional vehicle to operationalise sustainable data commons. She will conduct qualitative research for two case-studies of data commons: community-created language datasets, and voluntary pooling and donation of health data by individuals. Through these two case-studies, Ms. Chandrasekhar will build a blueprint for a data trust – teasing out the roles and obligations of the data trustee, the value of "trust" based on the Anglo-Saxon notion of public trust, and the techno-legal architecture of a data trust (including technical interfaces, new licenses, and other legal instruments). Ms. Chandrasekhar's doctoral thesis is intended to contribute to global data law and governance of collectives and is therefore not specific to a jurisdiction, but she will analyse jurisdiction-specific aspects of trust law as relevant for her case-studies.

II.   Nature of digital collaboration (1 page max)

Ms. Chandrasekhar's PhD research will investigate techno-legal collaborations for sustainable production, sharing and reuse of digital data.

First, Ms. Chandrasekhar is interested in the production and sharing of digital data by communities collaborating online. Here, she is particularly interested in two types of communities – linguistic communities and patient communities. How do actors with different types of expertise in each of these communities come together to use technologies of digitisation and platformatisation as well as legal frameworks of contract and copyright to create, pool and share their data for research and innovation? For instance, linguistic communities in Sub-Saharan Africa create and maintain machine-readable language datasets for training AI models to which they attach their own data reuse licenses. Patient communities in the European Union such as rare disease networks and citizen science collectives pool and share personal health data for medical research and innovation, while documenting legal conditions for reuse of this data in smart contracts.

Second, Ms. Chandrasekhar is interested in technology-mediated collaborations for reuse of such types of community-created data. In particular, she is interested in whether data trusts as a techno-legal institution can ensure non-extractive use of community-created language and health data. On the one hand, as a technical interface attached to the datasets, can data trusts monitor use of these datasets as well as monitor compliance with licenses and contracts created by the communities? And does this amount to a type of "digital rights management" for the public good/in the public interest? On the other hand, as a legal institution, can data trusts monitor compliance with licenses and contracts of these communities, as well as negotiate new practices of interoperability?

Third, recognising the limits of private law in a political economy of digital capitalism, Ms. Chandrasekhar is also interested in other types of public law reforms necessary to prevent extractive use of community-created datasets by large platform companies. There are two sides to this coin, that are of equal importance. First, public legal reforms to curtail and break-up infrastructural dominance of large platform companies. Interoperability and portability requirements under the European Union's Digital Markets Act and Data Act are an example. Second, the creation of necessary incentives for large platform companies to engage with and respect data trusts. Certain kinds of regulatory exemptions (such as exemptions under the AI Act for developers of free and openly licensed general-purpose AI models, who do not need to separately provide model documentation and information for downstream users) could act as a "carrot", while strategic litigation could act as the "stick". While this is not the central focus of her doctoral research, Ms. Chandrasekhar does seek to explore these aspects through academic papers and academic collaborations with peers (particularly peers in the PePR Ensemble community).

III.   Contribution to digital collaboration: Expected results and Impact (1 page max)

Ms. Chandrasekhar's doctoral research project aims to provide three sets of results.

First, through desk research, Ms. Chandrasekhar will make theoretical contributions on the role of law for sustainable data commons. While there has been significant empirical research on digital commons (including data and knowledge commons), there is a dearth of normative legal research on this topic. As a result, Ms. Chandrasekhar will build a conceptual framework to evaluate the impact of commodification of open data by law on sustainability of both the resource and the community. Using this conceptual framework, Ms. Chandrasekhar will also advance a normative framework of data trusts as a legal vehicle for data commons, that can challenge the commodification of open data and support sustainable collaboration.

Second, Ms. Chandrasekhar will conduct qualitative research as part of her case-studies, focused on two types of communities – linguistic communities who create and maintain machine-readable language datasets, and patient communities who pool and donate personal health data for research and innovation. Ms. Chandrasekhar will aim to produce empirical results on: (i) how new licenses create new reciprocity obligations for sustainability of both resource and community, and (ii) the potentiality of data trusts to ensure such reciprocity. These qualitative results are also relevant to law and policymakers within the European Union, given that community-created datasets for AI training and health data donation are topics of contemporary interest. This qualitative research will also inform the third projected result, outlined below.

Third, Ms. Chandrasekhar's doctoral research contributes to a collaboration between the Centre Internet et Société of the CNRS and the Open Knowledge Foundation known as SUDACO (or Sustainable Data Commons), PI for CNRS is her PhD advisor Mélanie Dulong de Rosnay. By virtue of this collaboration, Ms. Chandrasekhar seeks to disseminate her research findings to practitioners, civil society and communities within the Open Knowledge Foundation network. Ms. Chandrasekhar's research findings will also serve as the basis for a collaboration itself, where Ms. Chandrasekhar, CIS-CNRS and OKFN partner with computer scientists, data scientists and visual designers to build a beta version of a data trust to govern data commons collectively according to shared values as presented in this paper.

IV.   Positioning in the eNSEMBLE program (1/2 page max)

I believe that Ms. Chandrasekhar's doctoral research project fits closely within the objectives of the PEPR since digital and data commons are landmark of online collaboration, and as of PC5 TRAVERSE, in particular, I believe her research findings will be of unique value for Work Package 3: Address legal, ethical and philosophical issues.

In analysing commodification of open data by law, Ms. Chandrasekhar will analyse how copyright and privacy law produce new tensions for the production of open data as also for equitable use of open data. Given the entrenchment of digital capitalism, can copyright and privacy law still be "hacked" to enable more sharing and reuse of data while also limiting extractive use? How are linguistic communities and patient communities creatively tinkering with these legal frameworks to articulate new desires for reciprocity, to ensure sustainable use of their data?

In studying and proposing data trusts as a legal-institutional vehicle for communities to enforce their demands for reciprocity, Ms. Chandrasekhar is also interested in evaluating if data trustees can cultivate new practices of interoperability-*by-doing*. For instance, can data trustees provide ex-ante advice to reusers on how to combine differently licensed artefacts, which in turn can yield contextual interoperability?

Ms. Chandrasekhar's research also focusses on data commons, i.e. collective action for governance of digital data. This is a crucial type of technology-mediated collaboration, and would be a novel addition to PC5 and to the PePR-Ensemble as a whole. Ms. Chandrasekhar is interested in contributing to activities of the Ensemble community in this regard, by presenting her research at seminars as well as collaborating with peers in the community.