OECD STIP Hackathon
Open Data Access

Ern Chern Khor, Jisoo Sim, Taylor De Rosa
Team KAIST

Table of contents

- Research question

- Data

- Methods and Findings

  - Descriptive statistics from COMPASS data by policy instrument

  - Typologies through topic modeling

- Limitations

- Future research direction

Research Question

Can we characterize typologies of policies in support of making research data from publicly funded research openly accessible and reusable to the largest extent possible?

In particular, policies that:

- Foster and support open access by default to research data and other research-relevant digital objects from public funding

- Develop infrastructure and services to facilitate the accessibility of research data and other research-relevant digital objects from public funding within and across scientific domains and disciplines

Data

STIP Compass policy database

- Filtered on policies flagged as TH10: Open science and enhanced access to publications and research data (475 policies).

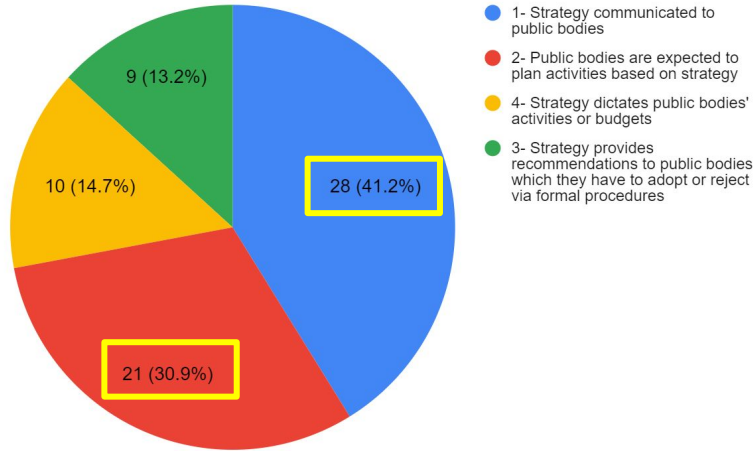- Gathered descriptive statistics across policy instruments within TH10-related policies.

TIP Strategies database

- Filtered on keywords = 'open_science', 'open_data', 'data_access', 'open_access', 'research_data', 'research_infrastructure', 'transparency'
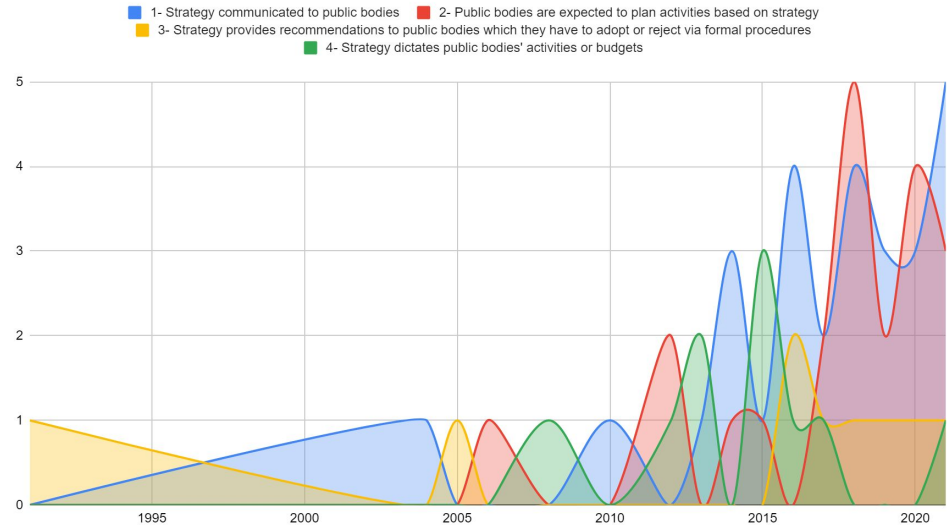
Findings – Descriptive statistics

Governance as policy instrument

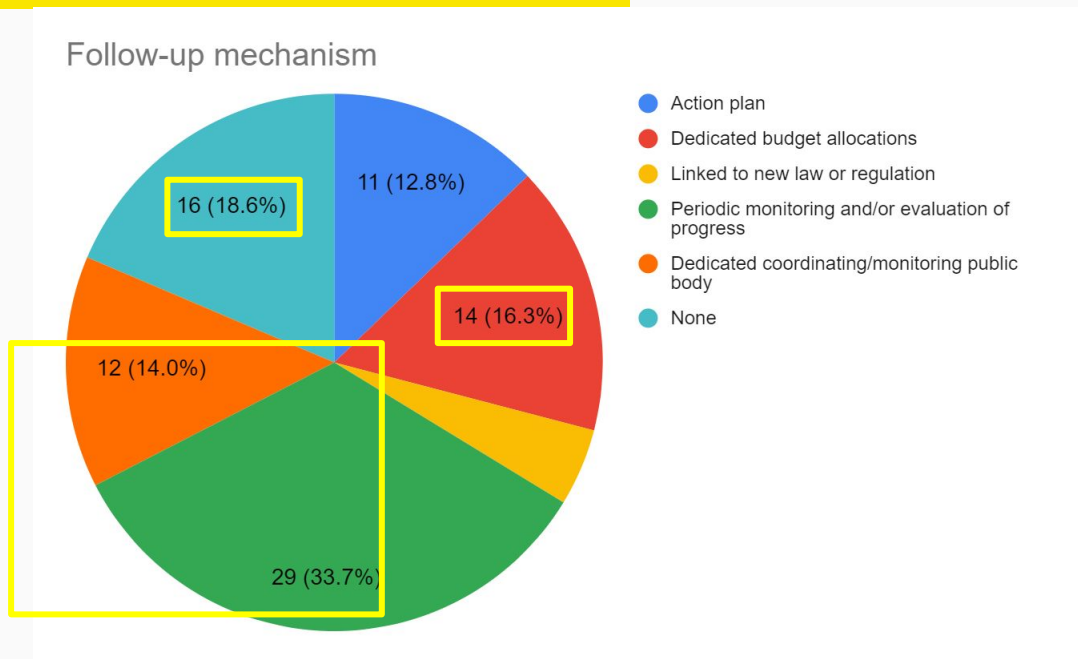Degree of coordination in implementing strategy

- 1- Strategy communicated to public bodies
- 2- Public bodies are expected to plan activities based on strategy
- 4- Strategy dictates public bodies' activities or budgets
- 3- Strategy provides recommendations to public bodies which they have to adopt or reject via formal procedures

28 (41.2%)
21 (30.9%)
10 (14.7%)
9 (13.2%)

Degree of coordination in implementing strategy by year

- 1- Strategy communicated to public bodies
- 2- Public bodies are expected to plan activities based on strategy
- 3- Strategy provides recommendations to public bodies which they have to adopt or reject via formal procedures
- 4- Strategy dictates public bodies' activities or budgets

Findings – Descriptive statistics

Governance as policy instrument

Follow-up mechanism

- Action plan
- Dedicated budget allocations
- Linked to new law or regulation
- Periodic monitoring and/or evaluation of progress
- Dedicated coordinating/monitoring public body
- None

11 (12.8%)

14 (16.3%)

16 (18.6%)

12 (14.0%)

29 (33.7%)

Findings – Descriptive statistics

Governance as policy instrument

Policy intelligence performed by
- Public administration
- Public research institute
- Academia
- Intergovernmental organisation
- Other

3 (23.1%)
3 (23.1%)
1 (7.7%)
1 (7.7%)
5 (38.5%)
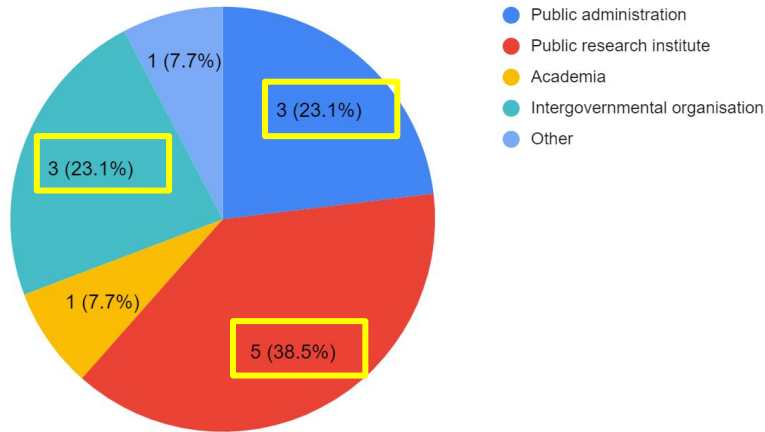
Policy intelligence performed by (by year)
- Public administration
- Public research institute
- Academia
- Private firms or consultants
- Civil society organisation
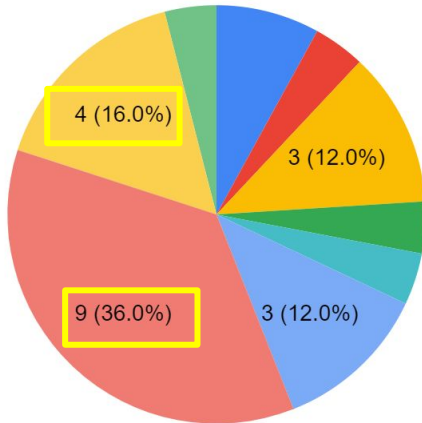- Intergovernmental organisation
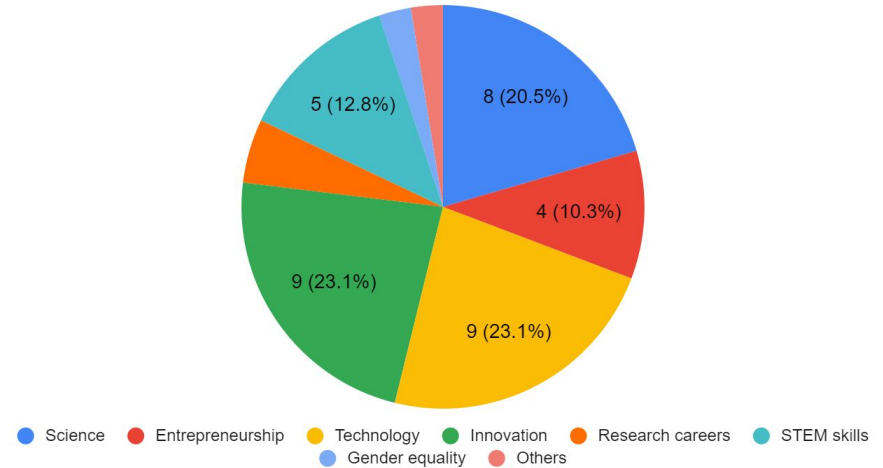- Other
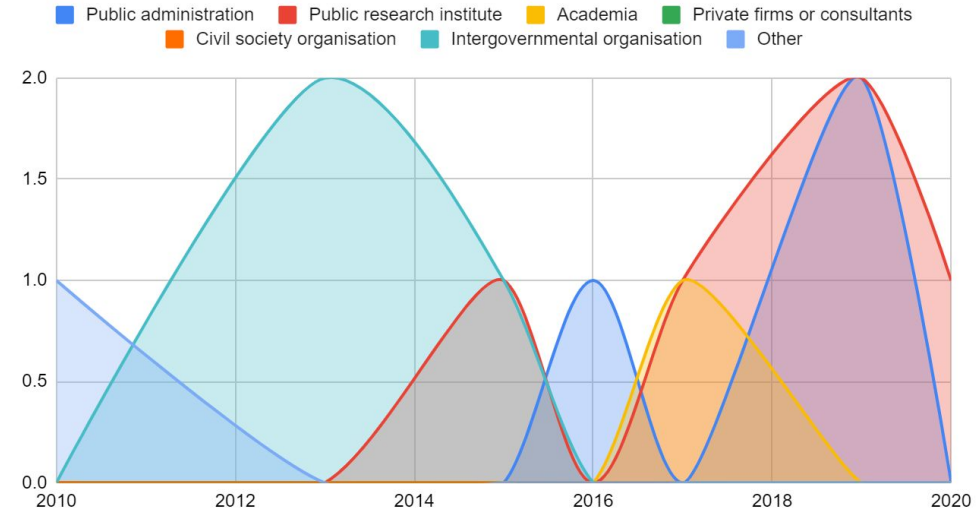
Findings – Descriptive statistics

Governance as policy instrument

Public awareness medium

- Public events
- School campaigns
- Conferences, workshops and/or training courses
- Museums
- Competitions
- Printed publications
- Websites
- Social media
- Science fairs

4 (16.0%)
3 (12.0%)
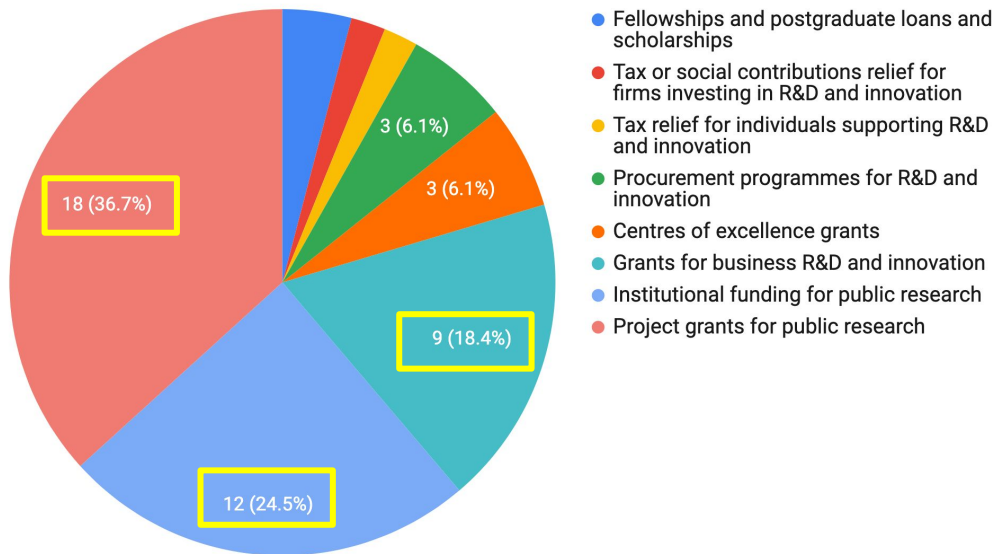9 (36.0%)
3 (12.0%)

Public awareness promoted..?

8 (20.5%)
4 (10.3%)
9 (23.1%)
9 (23.1%)
5 (12.8%)

- Science
- Entrepreneurship
- Technology
- Innovation
- Research careers
- STEM skills
- Gender equality
- Others

Findings - Descriptive statistics

Financial support as policy instrument
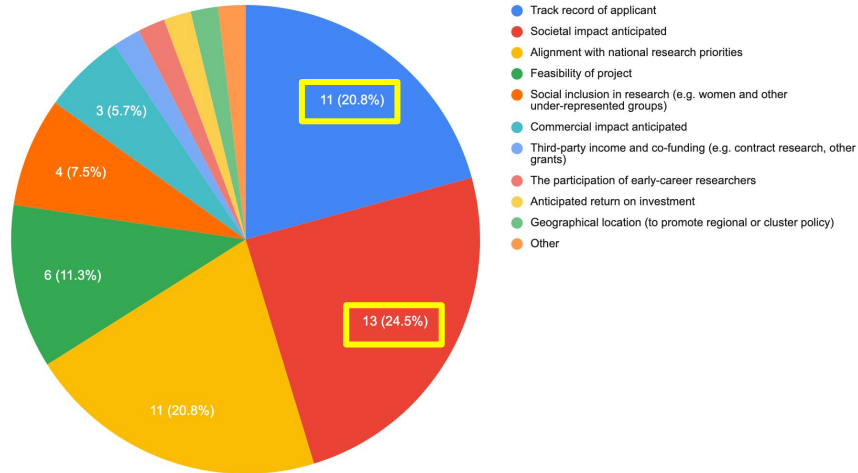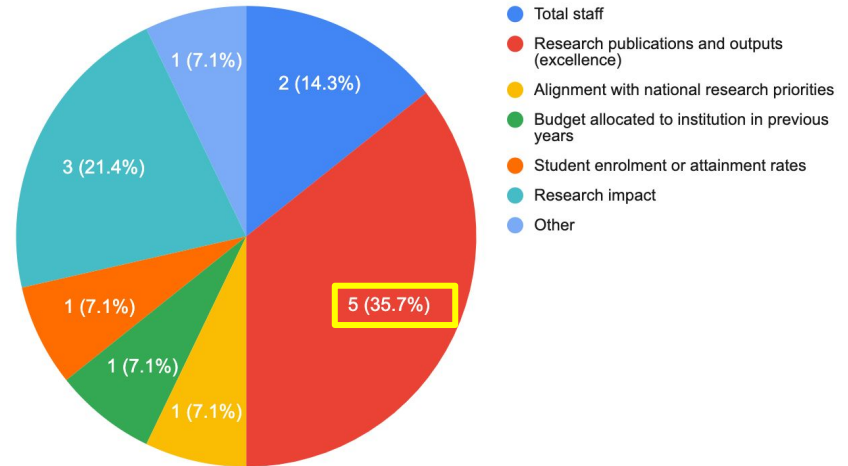
Findings – Descriptive statistics

Financial support as policy instrument – Selection Criteria

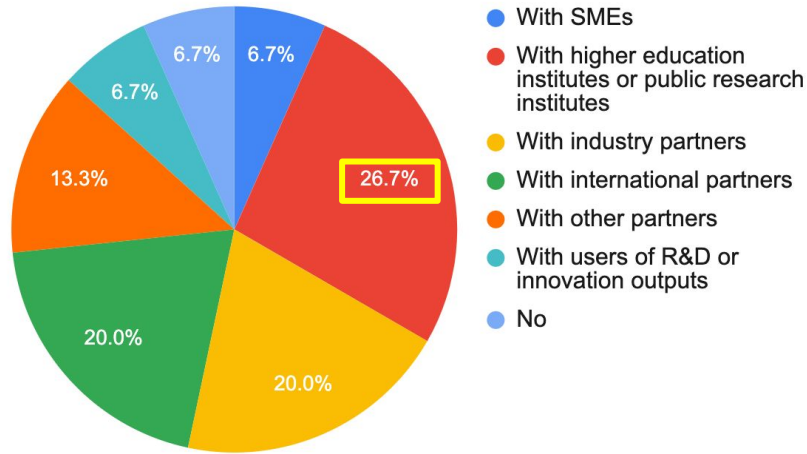**Selection Criteria - Project grants for public research & Grants for business R&D and innovation**

- Track record of applicant
- Societal impact anticipated
- Alignment with national research priorities
- Feasibility of project
- Social inclusion in research (e.g. women and other under-represented groups)
- Commercial impact anticipated
- Third-party income and co-funding (e.g. contract research, other grants)
- The participation of early-career researchers
- Anticipated return on investment
- Geographical location (to promote regional or cluster policy)
- Other

11 (20.8%) · 13 (24.5%) · 11 (20.8%) · 6 (11.3%) · 4 (7.5%) · 3 (5.7%)

Criteria for funding - Institutional funding for public research

- Total staff
- Research publications and outputs (excellence)
- Alignment with national research priorities
- Budget allocated to institution in previous years
- Student enrolment or attainment rates
- Research impact
- Other

2 (14.3%) · 5 (35.7%) · 3 (21.4%) · 1 (7.1%) · 1 (7.1%) · 1 (7.1%) · 1 (7.1%)
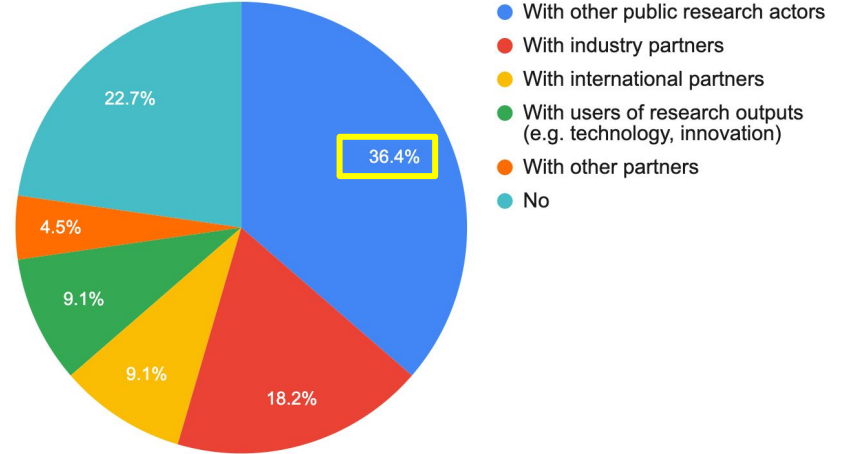
Findings – Descriptive statistics

Financial support as policy instrument – Requires collaboration with


Collaborators - Grants for business R&D and innovation

- With SMEs — 6.7%
- With higher education institutes or public research institutes — 26.7%
- With industry partners — 20.0%
- With international partners — 20.0%
- With other partners — 13.3%
- With users of R&D or innovation outputs — 6.7%
- No — 6.7%

Collaborators - Project grants for public research

- With other public research actors — 36.4%
- With industry partners — 18.2%
- With international partners — 9.1%
- With users of research outputs (e.g. technology, innovation) — 9.1%
- With other partners — 4.5%
- No — 22.7%

Findings – Descriptive statistics

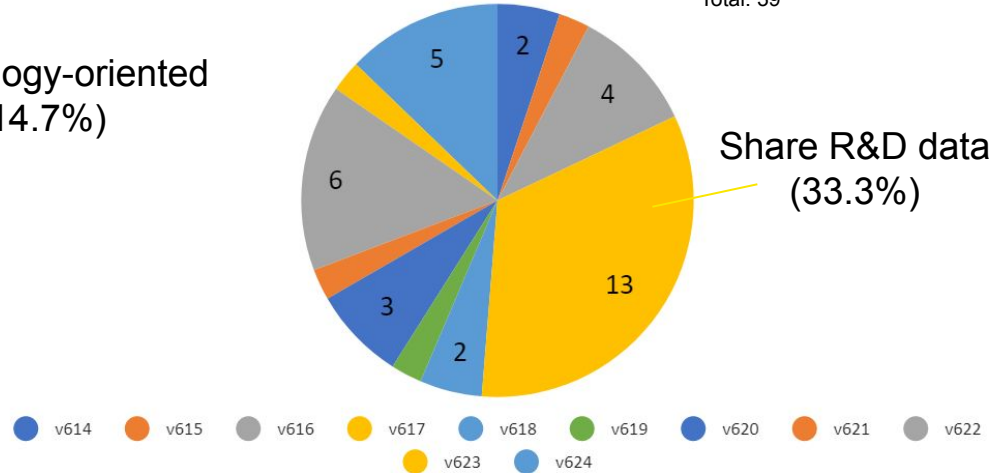## Collaborative infrastructure as policy instrument

Networking and collaborative platforms: Focus — Total: 34
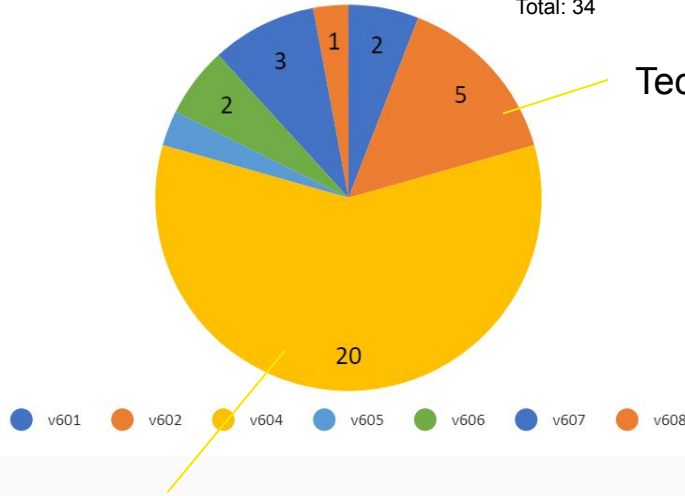
Technology-oriented (14.7%)

Research-oriented (58.8%)

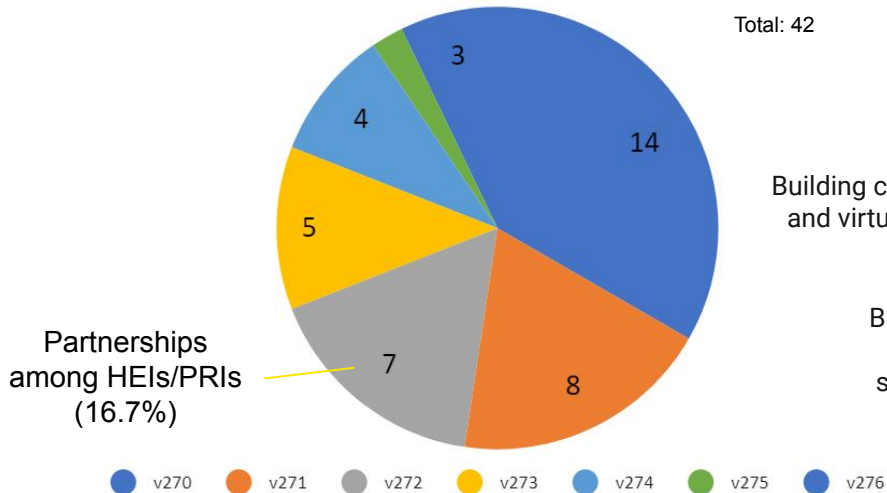Networking and collaborative platforms: Objective(s) — Total: 39

Share R&D data (33.3%)

Findings – Descriptive statistics

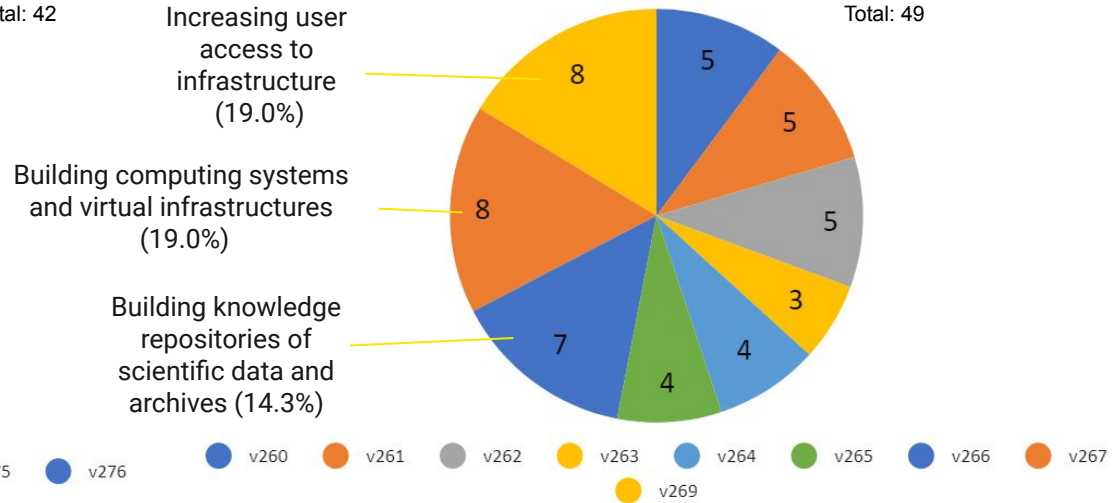## Collaborative infrastructure as policy instrument

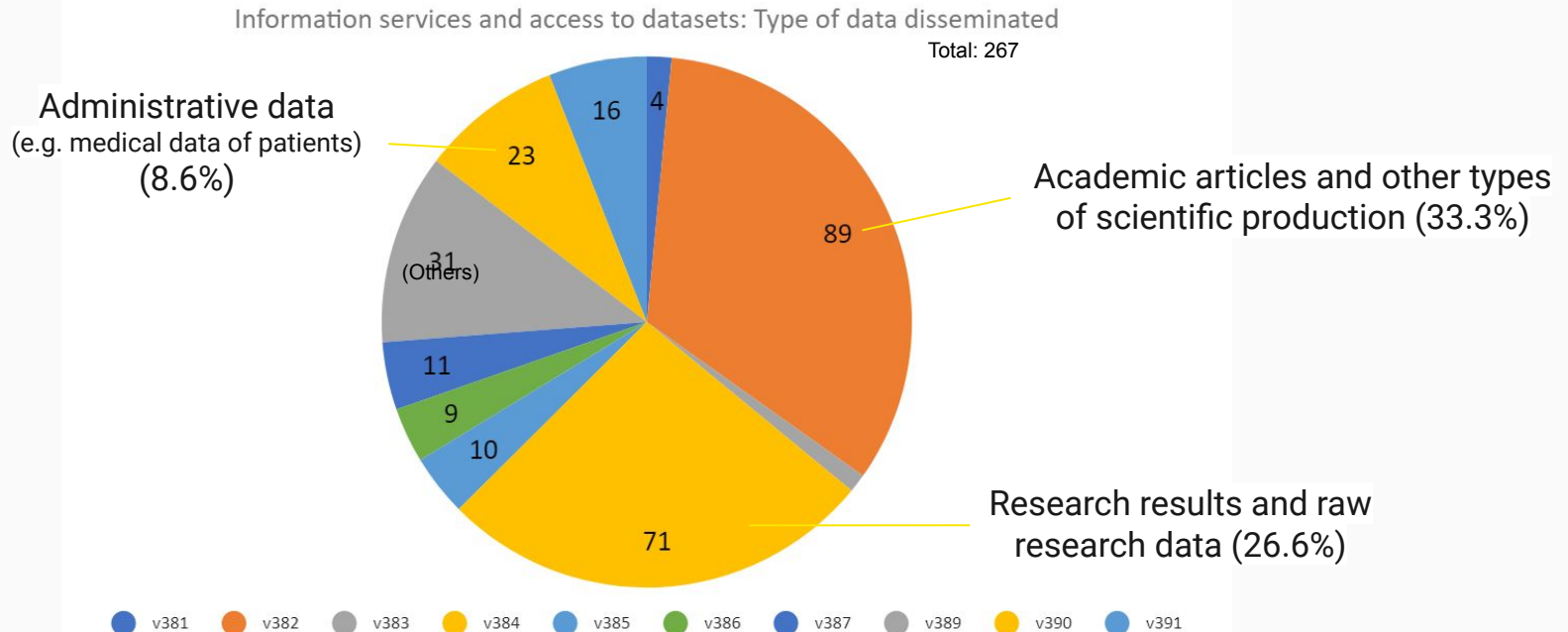Dedicated support to research infrastructures:Objective(s)

Total: 42

Increasing user access to infrastructure (19.0%)

Building computing systems and virtual infrastructures (19.0%)

Building knowledge repositories of scientific data and archives (14.3%)

Partnerships among HEIs/PRIs (16.7%)

Dedicated support to research infrastructures: Funding used

Total: 49

Legend (left chart): v270, v271, v272, v273, v274, v275, v276

Legend (right chart): v260, v261, v262, v263, v264, v265, v266, v267, v269

Findings – Descriptive statistics

Collaborative infrastructure as policy instrument

Information services and access to datasets: Type of data disseminated

Total: 267

Administrative data (e.g. medical data of patients) (8.6%)

Academic articles and other types of scientific production (33.3%)

(Others)

Research results and raw research data (26.6%)

89, 16, 4, 23, 31, 11, 9, 10, 71

v381 v382 v383 v384 v385 v386 v387 v389 v390 v391

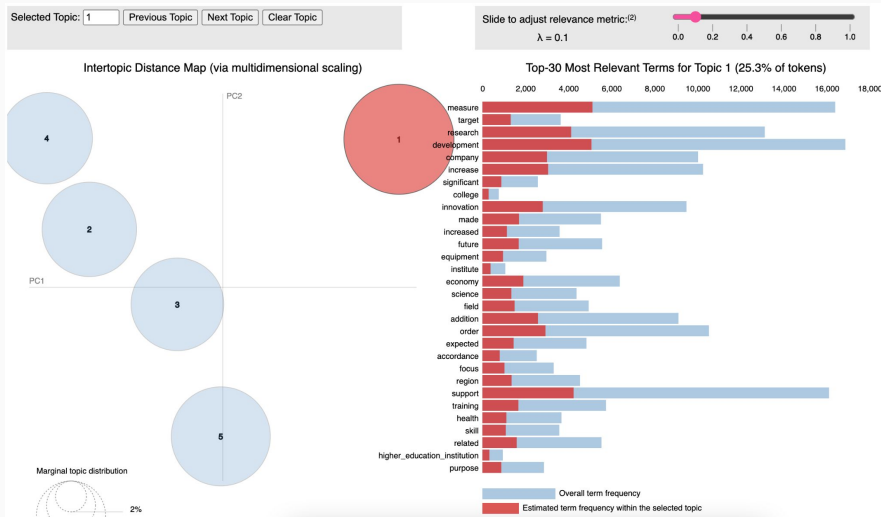# Findings – Typology using topic modeling

Data

- STIP Compass policy database: ShortDescription, Background, and Objectives(1-3) fields for each policy flagged as TH10.

- TIP Strategies database: clean_text for policies containing at least one of the keywords ('open_data', 'data_access', 'open_access', 'research_data', 'research_infrastructure', 'transparency')
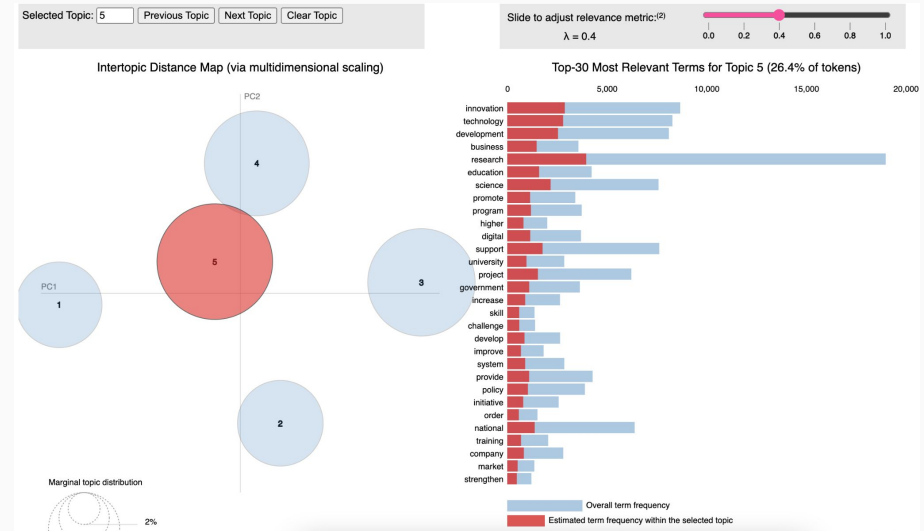
Gensim package – LDAMulticore model

Findings - Typology using topic modeling

TIP Strategies Policies

STIP Compass Policies

Findings – Typology using topic modeling

TIP Strategies Policies

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
College **Higher_education_institution** **Innovation** **Research** **Measure**	**Government** **Industry** **International** **Promoting** **Service**	**Digital** **Technology** **AI** **Network** **information**	**Environment** **CO2** **Bioeconomy** **Investment** **Project**	**Increase** **Reform** **Process** **Base** **Amount**

STIP Compass Policies

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Business **Investment** **Company** **Sector** **Innovation**	**Science** **International** **Cooperation** **Research** **Country**	**National** **Science** **Development** **Researcher** **Education**	**Public** **Energy** **Green** **Technology** **Health**	**Development** **Promote** **Support** **Research** **Education**

Limitations

- STIP Compass policy database provides categorical data, but these data are limited to understand each case in depth. The variables might have to be interpreted in different ways in different contexts.

- Both dataset focuses on policy description, but no data on policy evaluation, so insights are limited on real effect of policies.

- TIP Strategy database provides cleaned data with word tokens, but it is very challenging to tokenize text by sentences (which can be more insightful), due to inconsistent format in raw text.

Future direction

- Data quality
  - Reinforcing in-depth data which will complement the information from categorical data and enable better understanding to each policy
  - Resolving "missing values" and "everything looks similar" problem
  - Intermediary cleaned data of descriptive text
- Policy side
  - Research on outcomes (e.g. accessible repository status)
  - Encouraging international cooperation for research data access and sharing best relative policies

THANK YOU

OECD STIP Hackathon
Team KAIST