AN END-TO-END RESEARCH INFRASTRUCTURE FOR GENERATING AND SHARING DIVERSITY-AWARE DATA

DHHAI WORKSHOP

Munich, 26th June 2023

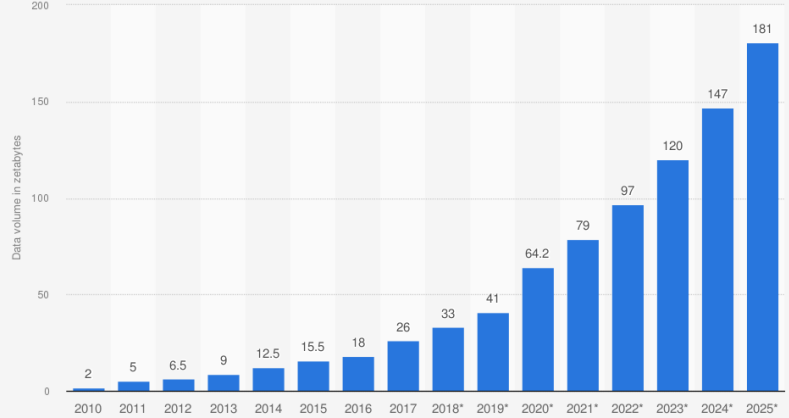
MATTEO BUSSO[1], Ronald Chenu[2] and Amalia de Götzen[2]

[1]University of Trento (Italy), [2]Aalborg University (Denmark)

Index

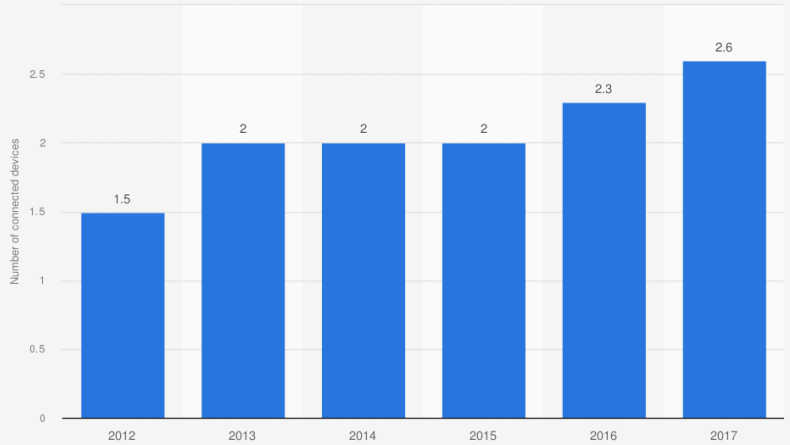
1. **Introduction**
2. Why a diversity aware RI?
3. What type of RI?
4. Current platforms
5. Benchmark
6. A former approach
7. Limits

A world of data



Sources
IDC; Seagate; Statista estimates
© Statista 2022

Additional Information:
Worldwide; 2010 to 2020



Sources
TNS Infratest; Google
© Statista 2022

Additional Information:
Italy; TNS Infratest; Google; 2012 to 2017; 1,000 respondents; 16 years and older; Computer-assisted telephone interview

1. Introduction
2. **Why a diversity aware RI?**
3. What type of RI?
4. Current platforms
5. Benchmark
6. A former approach
7. Limits

The computational turn

"[…] **digital technology is fundamentally changing** the way in which we engage in **the research process**. Many argue that this mediation is slowly beginning to change what it means to undertake research, affecting both the **epistemologies and ontologies** that underlie a research programme." (Berry, 2011)

Bias in data analysis and reuse

- **COMPAS:** white offenders were identified to be labeled as lower risk more likely than black offenders despite their criminal history
- **Amazon's hiring algorithm:** reinforces the gender gap in hiring
- **Reproducibility crisis:** the results of many scientific studies are difficult or impossible to reproduce

Privacy issues

Personal data in Google and Facebook (Asunciòn, 2017)	lack of valid consent given by their users
	insufficient access and control given to users over their personal information
	the risk of reidentification of anonymous personal data

1. Introduction
2. Why a diversity aware RI?
3. **What type of RI?**
4. Current platforms
5. Benchmark
6. A former approach
7. Limits

What is a Research Infrastructure?

According to EU Commission[1], RIs are ”facilities that provide resources and services for the research communities to conduct research and foster innovation in their fields”

[1] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/research-infrastructures_en

What resources?

COLLECTION

MANAGEMENT

DISTRIBUTION

COMMUNITY

1. Introduction
2. Why a diversity aware RI?
3. What type of RI?
4. **Current platforms**
5. Benchmark
6. A former approach
7. Limits

Data collection

Data from user interaction	• Psychlog • Mobile Sensing Platform
Sensor data	• ResearchStack
Interaction and sensors	• AWARE • CS Logger

Data management and distribution

UkDataArchive

GESIS

Open Science Framework (OSF)

ComputerOntario

BioBank

Crowd-sourcing vs. Communities

Pervasiveness of smart devices to collect data on large panel, e.g.:

- Participatory sensing
- Mobile Crowd Sensing

Limits

- Western countries panel
- Participant as a sensor

Focus on involvement and education, e.g.:

- Natural science: Zooniverse, CornellLab, iNaturalist
- Broader focus: SciStarter, EU citizen-science

Limits:

- Missing diversity aware data

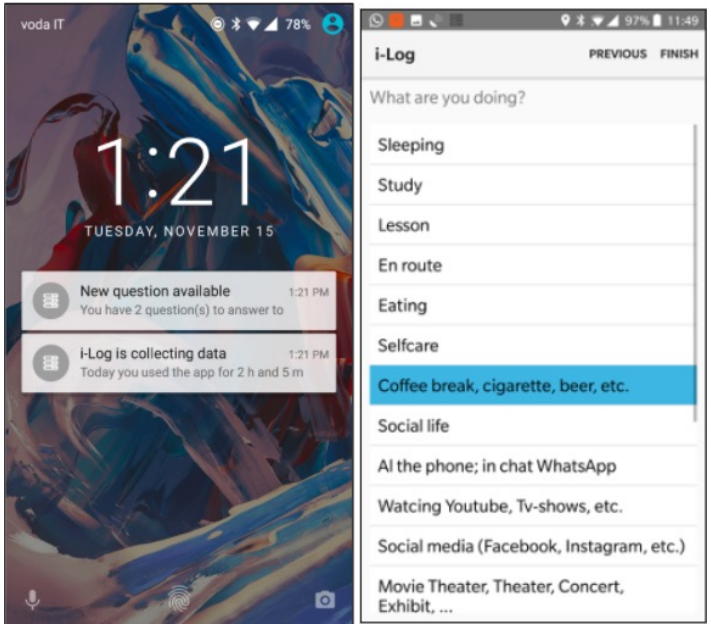
1. Introduction
2. Why a diversity aware RI?
3. What type of RI?
4. Current platforms
5. **Benchmark**
6. A former approach
7. Limits

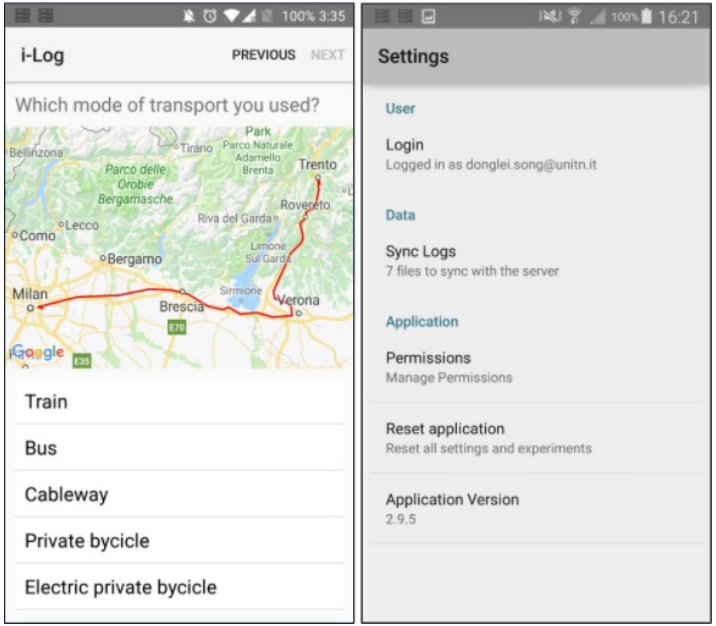
Benchmark

	Collection	Ethics & GDPR	Management	Distribution	Community
AWARE	✓		✓		
Citizen Science Logger	✓		✓		✓
UK Data Archive		✓	✓		✓
Zooniverse	✓			✓	✓

1. Introduction
2. Why a diversity aware RI?
3. What type of RI?
4. Current platforms
5. Benchmark
6. **A former approach**
7. Limits

iLog App





Design: Support material

Design & Management

- Research design and management
- Research process Gantt
- iLog questions and sensors specifications

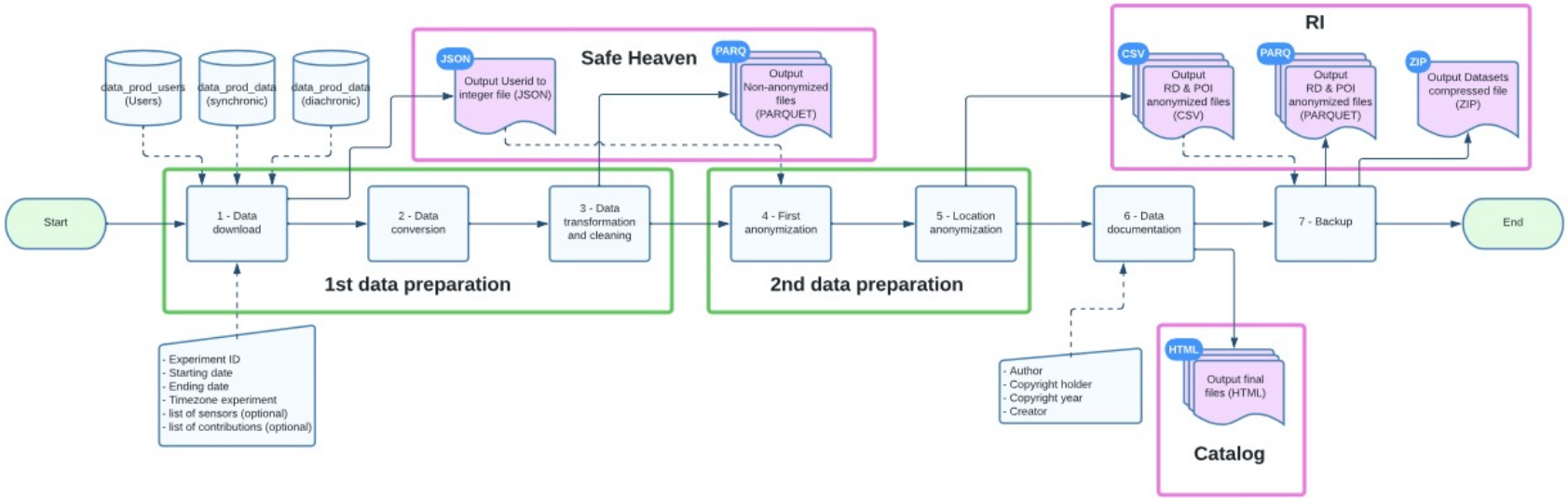
GDPR & Ethics

- Informed consent
- Privacy statement
- DPIA (Processor and Controller)
- Ethical committee approval

Guidelines

- Ethics Self Assessment Tool – UK Statistics Authority
- Hands on guide to research

Data preparation



contributionanswers

Overview Variables Datasets

Contributionanswers

Questionnaires

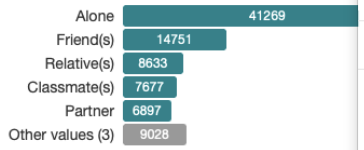
Airplane mode event
Application event
Battery monitoring log
Batterycharge event
Cellular network
Doze event
Headset plug event
Location event per time POI
Location event per time RD
Music event
Notification event
Proximity event
Ring mode event
Screen event
Wifi event
Wifi networks event

A5
Categorical

MISSING

With whom are you?

Distinct	8
Distinct (%)	< 0.1%
Missing	79840
Missing (%)	47.5%
Memory size	1.3 MiB



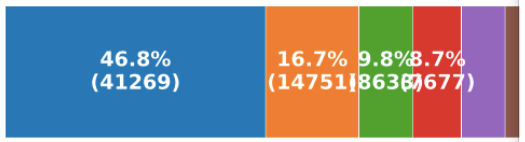
Toggle

Overview Categories

Common Values

Value	Count	Frequency (%)
Alone	41269	24.6%
Friend(s)	14751	8.8%
Relative(s)	8633	5.1%
Classmate(s)	7677	4.6%
Partner	6897	4.1%
Roommate(s)	6265	3.7%
Colleague(s)	1470	0.9%
Other	1293	0.8%
(Missing)	79840	47.5%

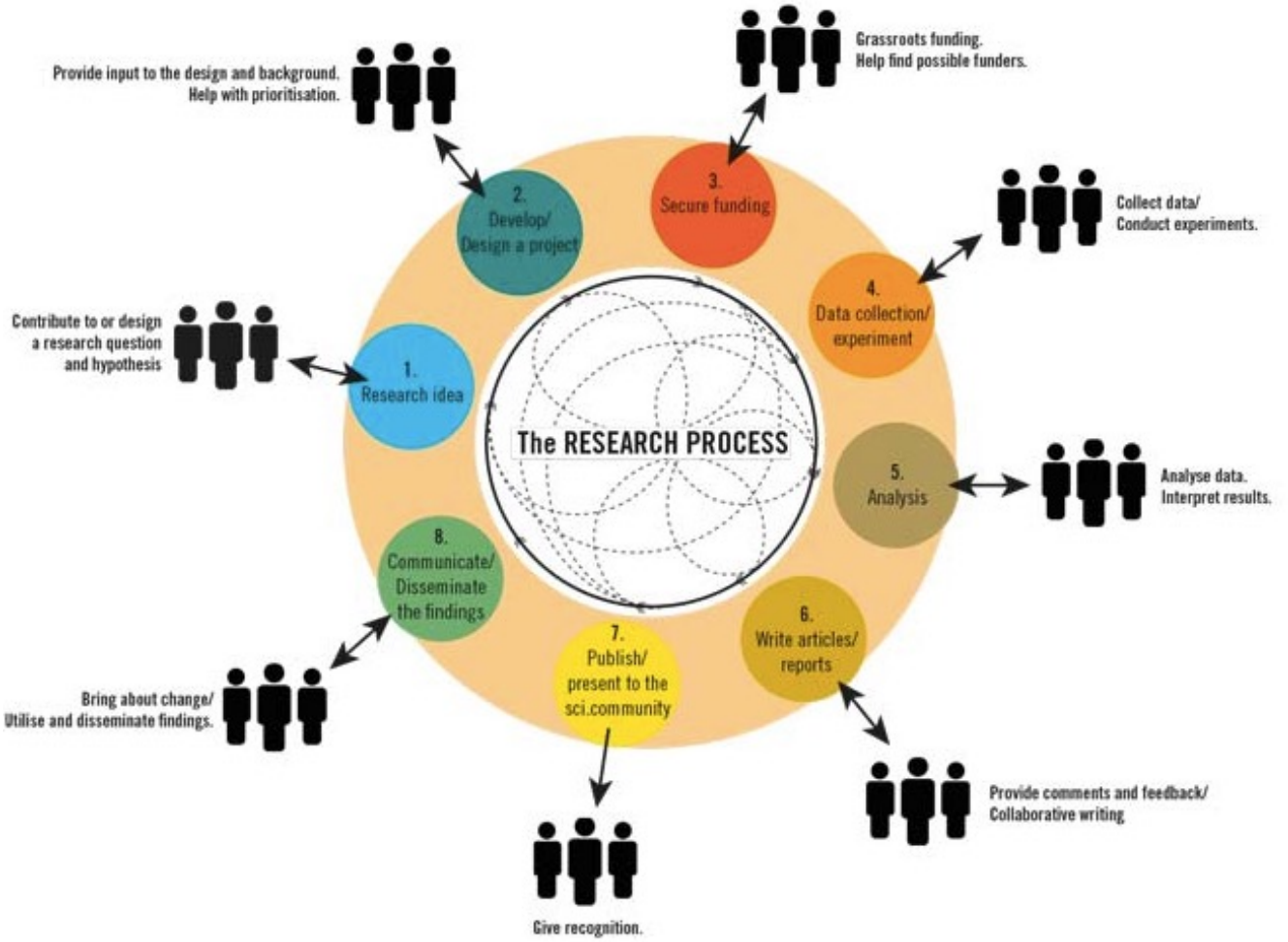
Category Frequency Plot



The LivePeople catalog

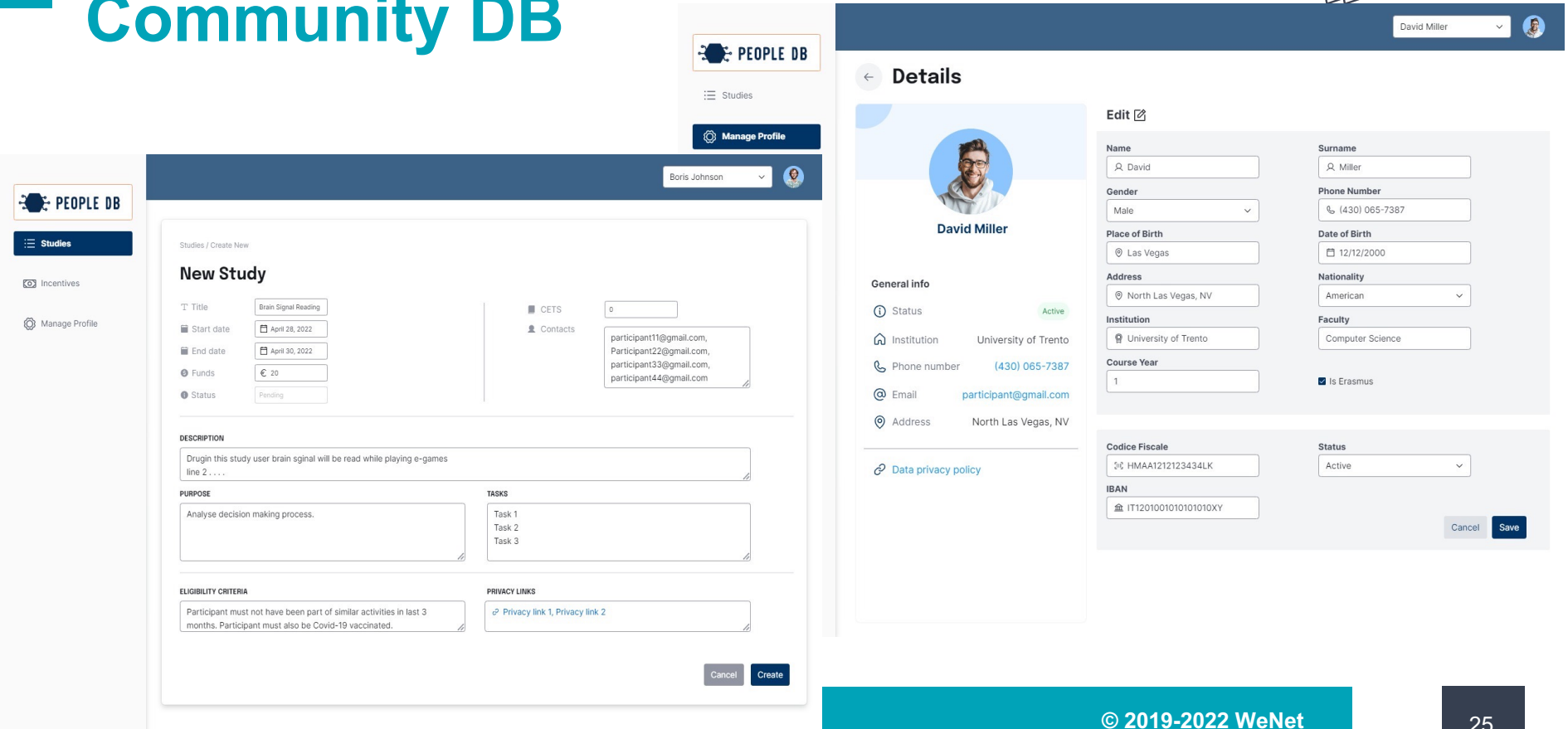
Community

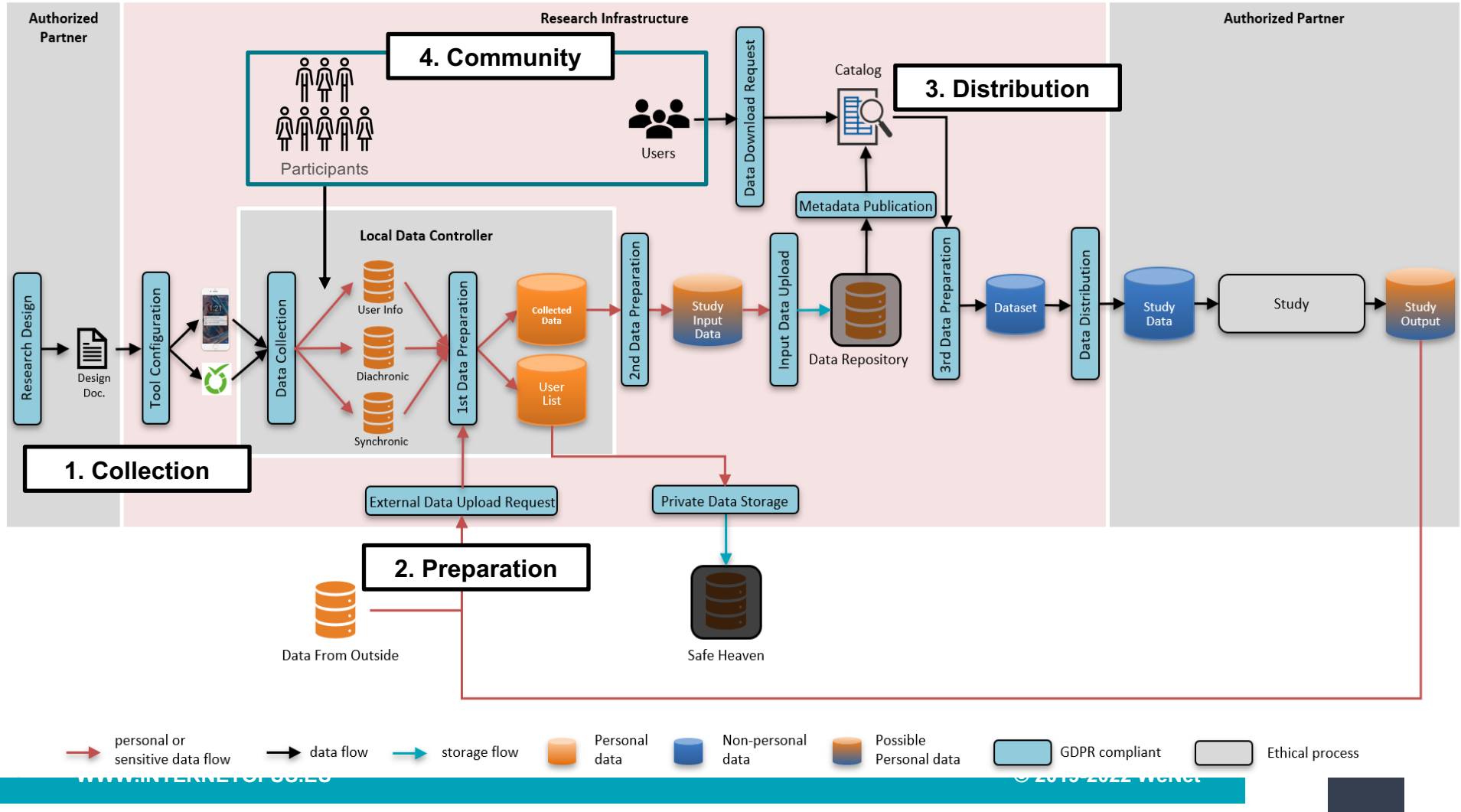
Generating spaces for interactions:
among participants and researcher (e.g., between students and professors)
within participants or researcher (e.g., sharing knowledge on study design)



VA (Public & Science): https://v-a.se/english-portal/citizen-science/

Community DB




Authorized Partner
Research Infrastructure
Authorized Partner
4. Community
Participants
Users
Data Download Request
Catalog
3. Distribution
Metadata Publication
Local Data Controller
Research Design
Design Doc.
Tool Configuration
Data Collection
User Info
Diachronic
Synchronic
1st Data Preparation
Collected Data
User List
2nd Data Preparation
Study Input Data
Input Data Upload
Data Repository
3rd Data Preparation
Dataset
Data Distribution
Study Data
Study
Study Output
1. Collection
External Data Upload Request
Private Data Storage
2. Preparation
Data From Outside
Safe Heaven
personal or sensitive data flow
data flow
storage flow
Personal data
Non-personal data
Possible Personal data
GDPR compliant
Ethical process

1. Introduction
2. Why a diversity aware RI?
3. What type of RI?
4. Current platforms
5. Benchmark
6. A former approach
7. **Limits**

Limits

(i) the usability of iLog app by non-expert users, such as citizens;

(ii) the validation of data management outcomes to effective reuse of resources;

(iii) the community is still being tested and non-monetary incentive strategy is missing

GET IN TOUCH

http://knowdive.disi.unitn.it/

http://datascientia.disi.unitn.it/

@knowdive

matteo.busso@unitn.it

THANK YOU!

WeNet project is funded by the EU's Horizon2020 programme under Grant Agreement number 823783.