LUND UNIVERSITY

# Open data ecosystems – wishful thinking or successful business?

PROFESSOR PER RUNESON ✉ PER.RUNESON@CS.LTH.SE @SOFTENGRESGRP

# Data science

…is an **interdisciplinary** field that uses **scientific methods**, processes, algorithms and systems to **extract knowledge** and insights from noisy, **structured and unstructured data**, and apply knowledge and actionable insights from data across a broad range of **application domains**. [Wikipedia]

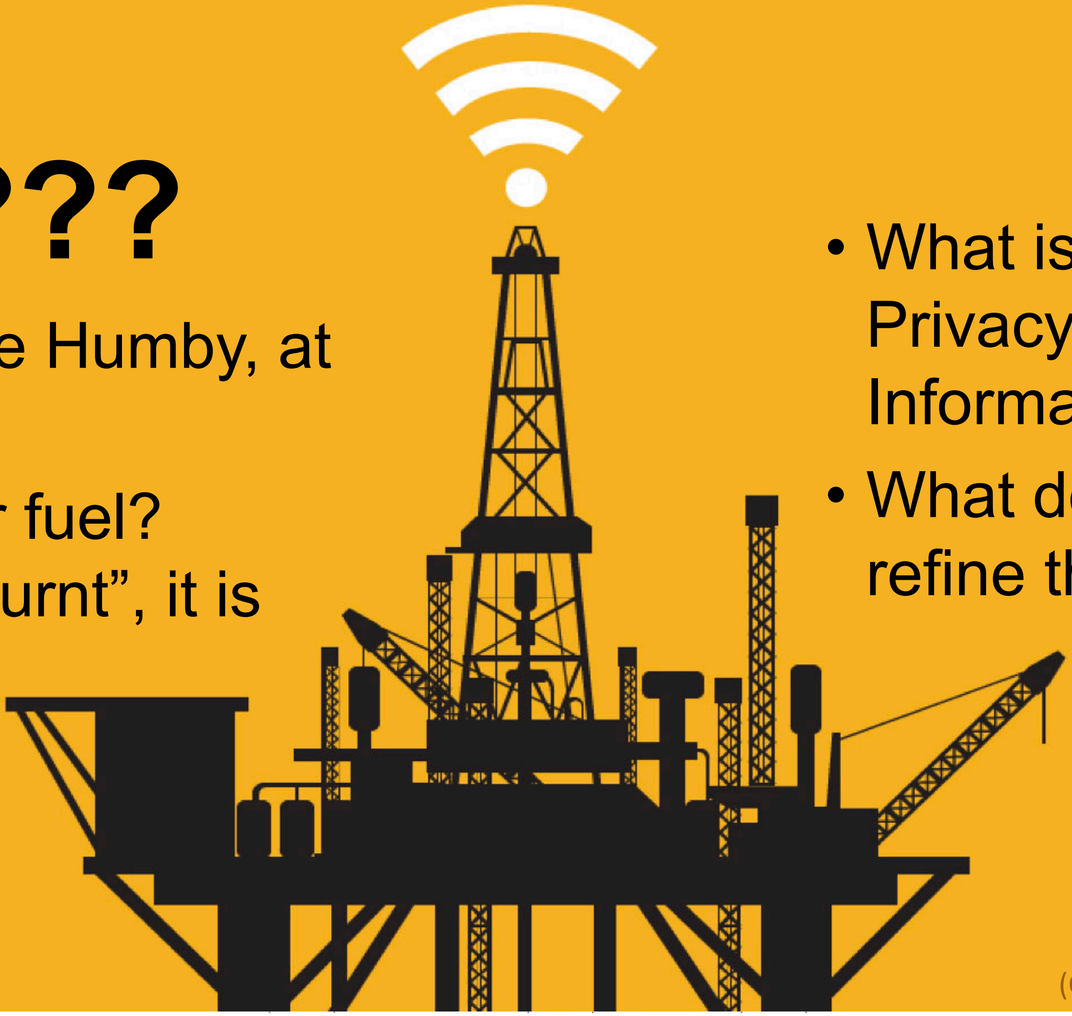
But where does the data come from?

Do you have enough quality data?

Can you afford maintaining the data?

| Date | Post | Reach | Engaged Users | Talking About This | Virality |
|---|---|---|---|---|---|
| 10/07/11 | "Don't spend more time plannin... | 1,878 | 42 | 12 | 0.64% |
| 10/06/11 | http://www.bethkanter.org/bt... | 1,733 | 20 | 6 | 0.35% |
| 10/06/11 | What do you think about the n... | 1,934 | 70 | 22 | 1.14% |
| 10/05/11 | Flash Poll: If you are a nonprofi... | 1,988 | 94 | 9 | |
| 10/05/11 | Just noticed that the new metr... | 2,006 | | | |
| 10/05/11 | If you are | | | | |

# Data is the new oil!



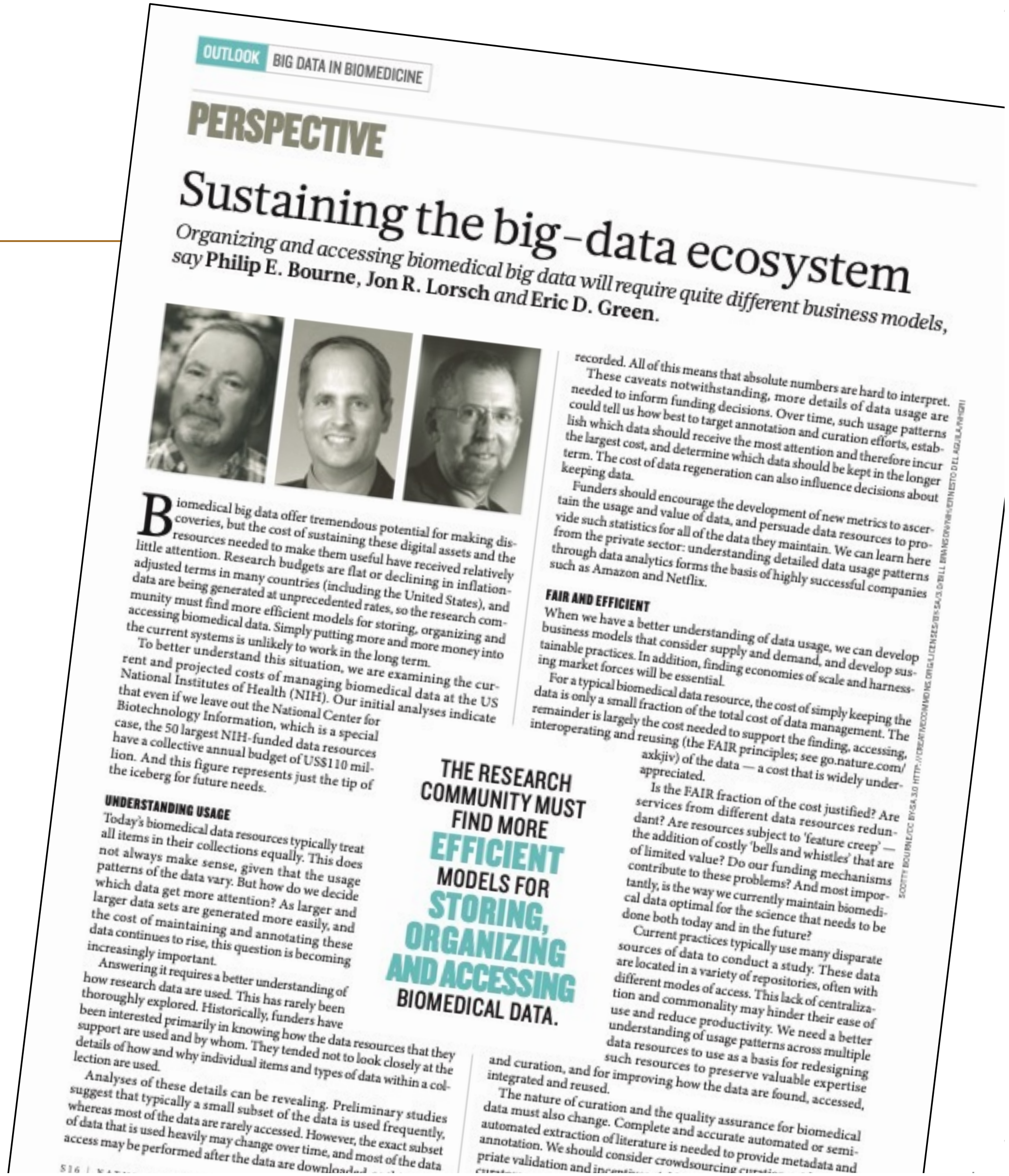
## Is it ????

- Claim by Clive Humby, at Tesco, 2006
- Lubrication or fuel? Data is not “burnt”, it is non-rival

- What is data pollution? Privacy intrusion? Information leakage?
- What does it cost to refine the oil?

# Example biomedicine

”For a typical biomedical data resource, the cost of simply keeping the data is only a small fraction of the total cost of data management. The remainder is largely the cost needed to support the finding, accessing, interoperating and reusing of the data — a cost that is widely under-appreciated.”



OUTLOOK BIG DATA IN BIOMEDICINE

PERSPECTIVE

## Sustaining the big-data ecosystem

*Organizing and accessing biomedical big data will require quite different business models, say* **Philip E. Bourne, Jon R. Lorsch** *and* **Eric D. Green.**

Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-adjusted terms in many countries (including the United States), and data are being generated at unprecedented rates, so the research community must find more efficient models for storing, organizing and accessing biomedical data. Simply putting more and more money into the current systems is unlikely to work in the long term.

To better understand this situation, we are examining the current and projected costs of managing biomedical data at the US National Institutes of Health (NIH). Our initial analyses indicate that even if we leave out the National Center for Biotechnology Information, which is a special case, the 50 largest NIH-funded data resources have a collective annual budget of US$110 million. And this figure represents just the tip of the iceberg for future needs.

THE RESEARCH COMMUNITY MUST FIND MORE EFFICIENT MODELS FOR STORING, ORGANIZING AND ACCESSING BIOMEDICAL DATA.

### UNDERSTANDING USAGE

Today's biomedical data resources typically treat all items in their collections equally. This does not always make sense, given that the usage patterns of the data vary. But how do we decide which data get more attention? As larger and larger data sets are generated more easily, and the cost of maintaining and annotating these data continues to rise, this question is becoming increasingly important.

Answering it requires a better understanding of how research data are used. This has rarely been thoroughly explored. Historically, funders have been interested primarily in knowing how the data resources that they support are used and by whom. They tended not to look closely at the details of how and why individual items and types of data within a collection are used.

Analyses of these details can be revealing. Preliminary studies suggest that typically a small subset of the data is used frequently, whereas most of the data are rarely accessed. However, the exact subset of data that is used heavily may change over time, and most of the data access may be performed after the data are downloaded

recorded. All of this means that absolute numbers are hard to interpret.

These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in the longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies such as Amazon and Netflix.

### FAIR AND EFFICIENT

When we have a better understanding of data usage, we can develop business models that consider supply and demand, and develop sustainable practices. In addition, finding economies of scale and harnessing market forces will be essential.

For a typical biomedical data resource, the cost of simply keeping the data is only a small fraction of the total cost of data management. The remainder is largely the cost needed to support the finding, accessing, interoperating and reusing (the FAIR principles; see go.nature.com/axkjiv) of the data — a cost that is widely under-appreciated.

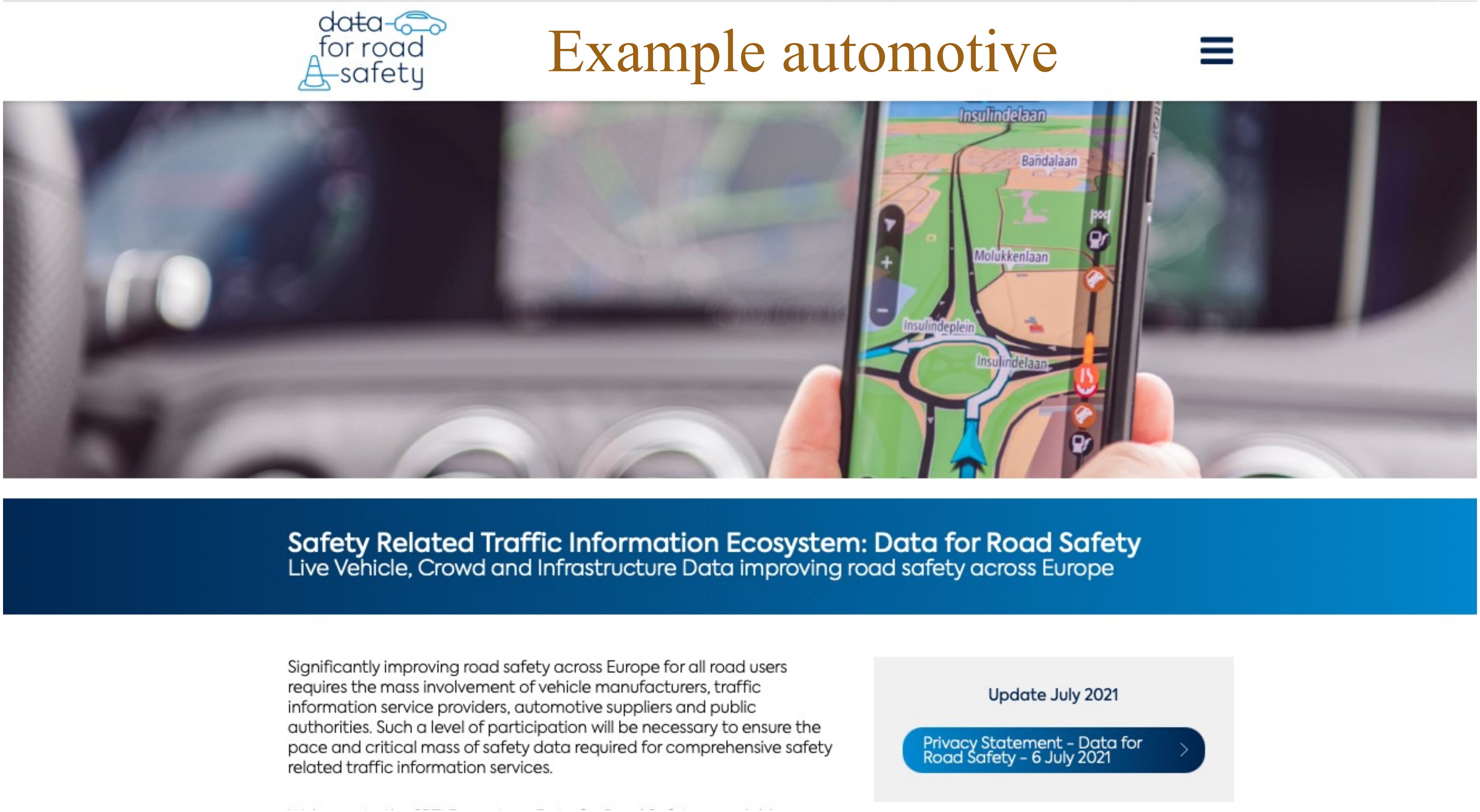
Is the FAIR fraction of the cost justified? Are services from different data resources redundant? Are resources subject to 'feature creep' — the addition of costly 'bells and whistles' that are of limited value? Do our funding mechanisms contribute to these problems? And most importantly, is the way we currently maintain biomedical data optimal for the science that needs to be done both today and in the future?

Current practices typically use many disparate sources of data to conduct a study. These data are located in a variety of repositories, often with different modes of access. This lack of centralization and commonality may hinder their ease of use and reduce productivity. We need a better understanding of usage patterns across multiple data resources to use as a basis for redesigning such resources to preserve valuable expertise and curation, and for improving how the data are found, accessed, integrated and reused.

The nature of curation and the quality assurance for biomedical data must also change. Complete and accurate automated or semi-automated extraction of literature is needed to provide metadata and annotation. We should consider crowdsourcing curation
priate validation and incenti

SCOTTY BOURNE/CC BY-SA 3.0 HTTP://CREATIVECOMMONS.ORG/LICENSES/BY-SA/3.0/BILL BRANSON/NIH/ERNESTO DEL AGUILA/NHGRI

# Example automotive



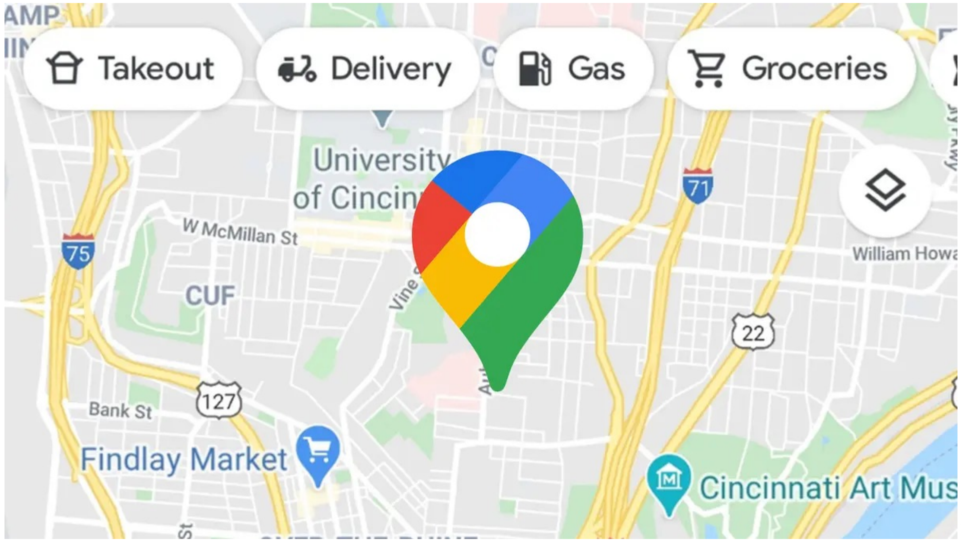
**Safety Related Traffic Information Ecosystem: Data for Road Safety**
Live Vehicle, Crowd and Infrastructure Data improving road safety across Europe

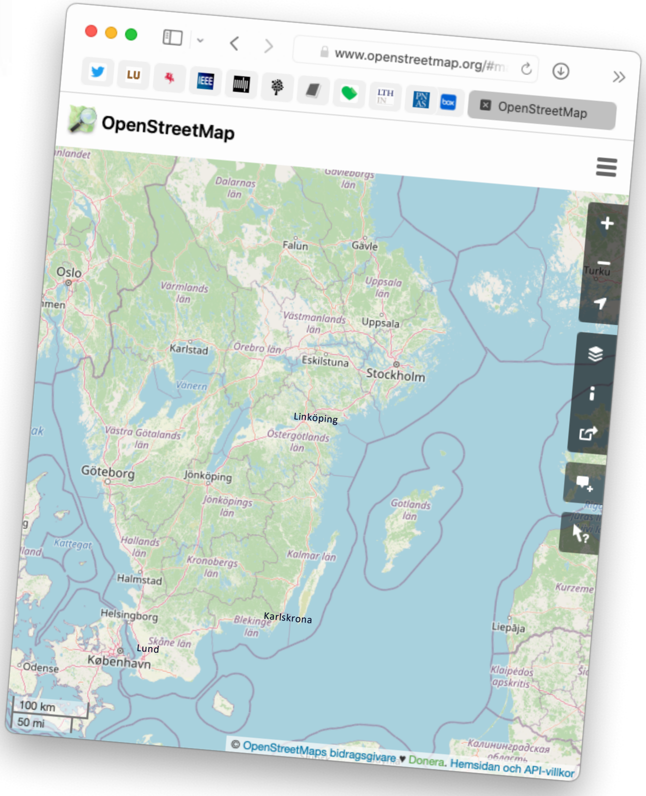
Significantly improving road safety across Europe for all road users requires the mass involvement of vehicle manufacturers, traffic information service providers, automotive suppliers and public authorities. Such a level of participation will be necessary to ensure the pace and critical mass of safety data required for comprehensive safety related traffic information services.

Update July 2021

Privacy Statement - Data for Road Safety - 6 July 2021 ›

# Example maps



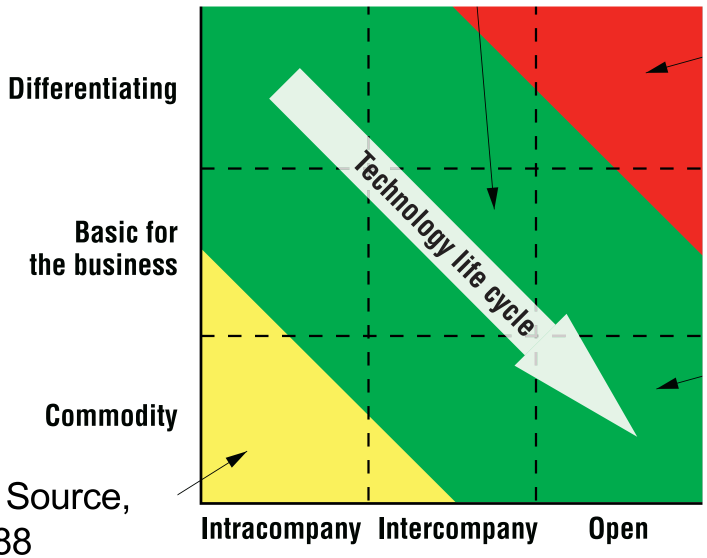


LUND UNIVERSITY

# Data challenges and opportunities

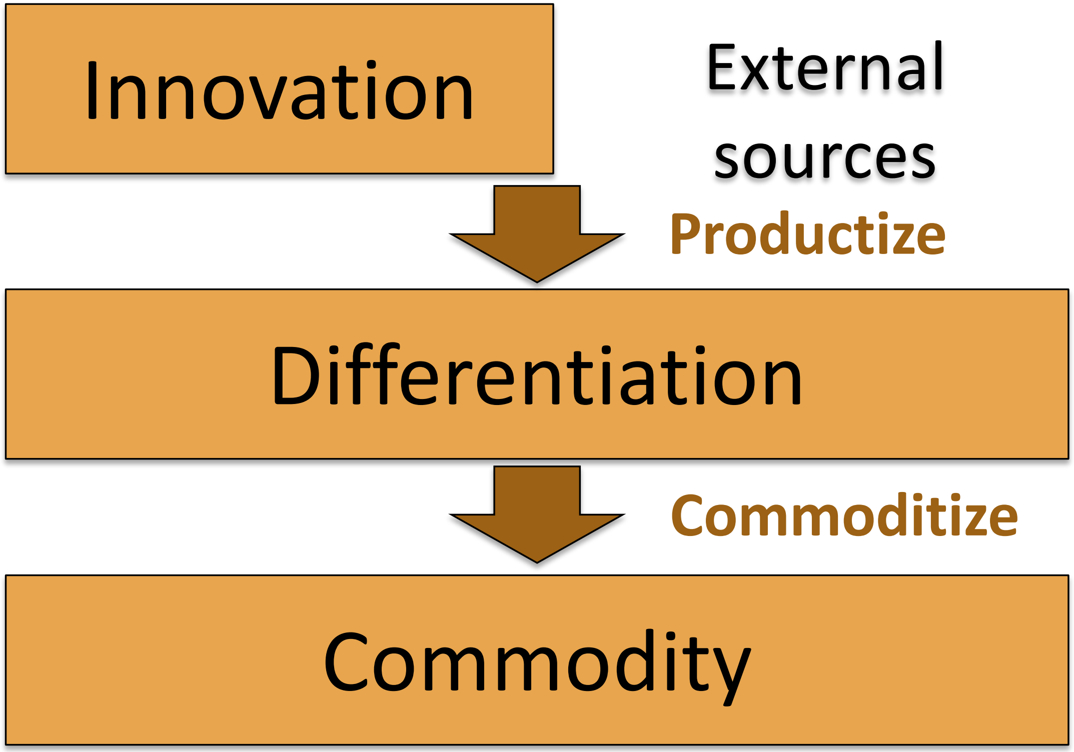
- Costs for *data maintenance*, *quality assurance* and *annotation* is a challenge
- Data will gradually become *commodity* for some functionality

Open data ecosystems?



Lundell *et al.* Commodification of Industrial Software: A Case for Open Source, *IEEE Software*, 26(04):77-83, 2009. https://doi.org/10.1109/MS.2009.88

# What is unique?



Innovation

External sources

Productize

Differentiation

Commoditize

Commodity

Achieving Simplicity with the Three-Layer Product Model

Jan Bosch, Chalmers University of Technology, Sweden

Novelty focus

Value focus

Cost focus

# Data sharing?

“Value comes from data being brought together, and that requires organizations to let others use the data they hold”

Bennett Institute for Public Policy
Cambridge

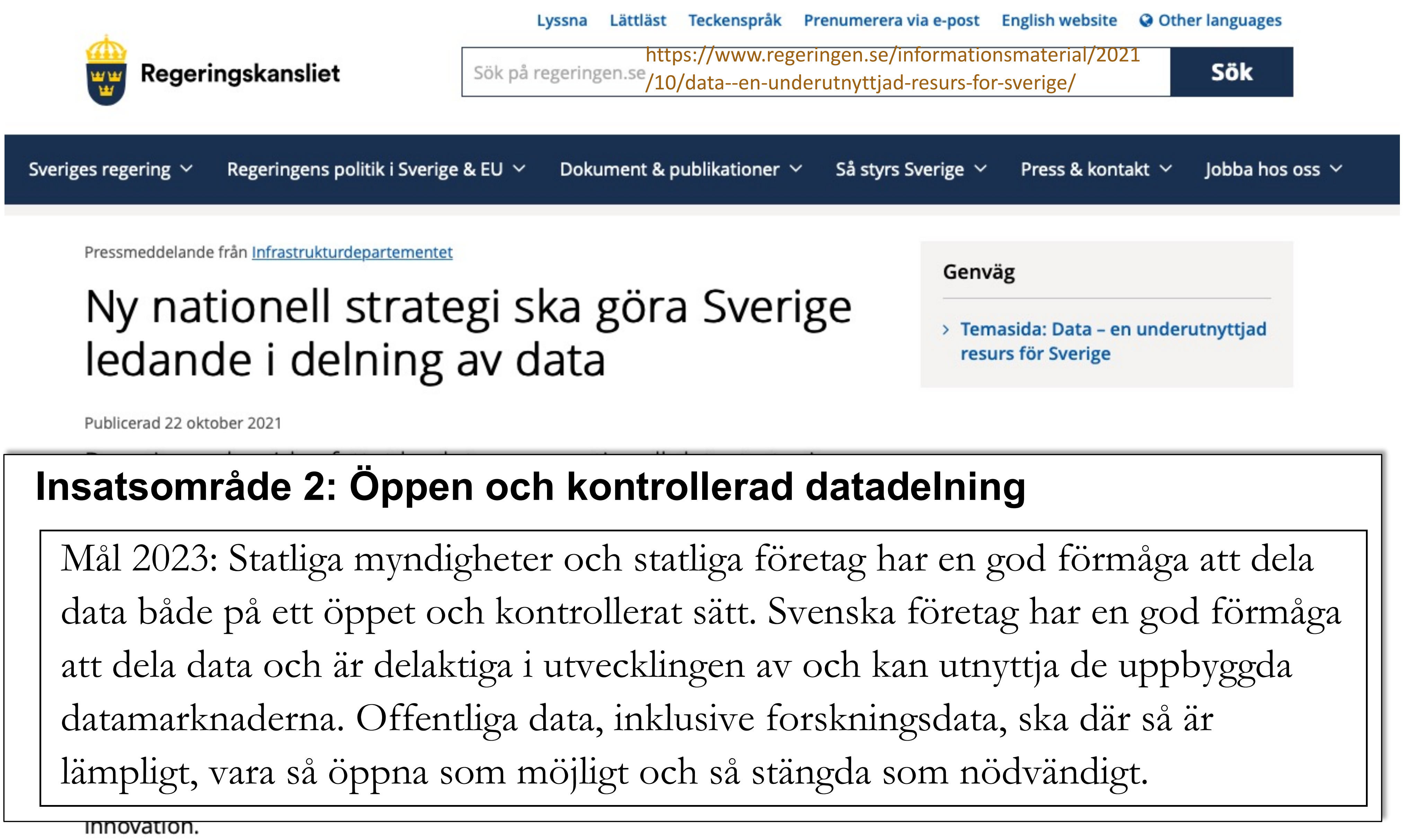
ODI

Nuffield Foundation

The value of data

The Value of Data summary report / 2020

https://www.bennettinstitute.cam.ac.uk/publications/value-data-summary-report/



Pressmeddelande från Infrastrukturdepartementet

# Ny nationell strategi ska göra Sverige ledande i delning av data

Publicerad 22 oktober 2021

innovation.

Genväg

› Temasida: Data – en underutnyttjad resurs för Sverige

https://www.regeringen.se/informationsmaterial/2021/10/data--en-underutnyttjad-resurs-for-sverige/

## Insatsområde 2: Öppen och kontrollerad datadelning

Mål 2023: Statliga myndigheter och statliga företag har en god förmåga att dela data både på ett öppet och kontrollerat sätt. Svenska företag har en god förmåga att dela data och är delaktiga i utvecklingen av och kan utnyttja de uppbyggda datamarknaderna. Offentliga data, inklusive forskningsdata, ska där så är lämpligt, vara så öppna som möjligt och så stängda som nödvändigt.

# Background and motivation
# Open source software practices

# Background – Open Source Software

- 1960/70's – software into the bargain
- 1980's – political movement
- 1990's – commercial (Linux)
- 2000's – databases (MySQL), Android
- 2010's – everywhere

OPEN SOURCE EXPANDED

EDITOR DIRK RIEHLE
Friedrich Alexander-University of Erlangen Nürnberg;
dirk.riehle@fau.de

## A Brief History of Free, Open Source Software and Its Communities

**Jesus M. Gonzalez-Barahona,** Universidad Rey Juan Carlos

*Free, open source software (FOSS) has a long history, beginning with the origins of software itself, when the terms free software and open source software were not yet defined. Learning about the milestones of this history may help to understand FOSS today*

computers, with IBM being, by a large margin, the market leader. For all of them, software was just a companion to hardware: as long as you paid for maintenance, you had access to the software catalog of the manufacturer. User groups, such as SHARE (IBM) and the DECUS [Digital Equipment Corp. (DEC)] favored software sharing. To some

to 1970

https://dx.doi.org/10.1109/MC.2020.3041887

# Open source in mobile devices – 2011

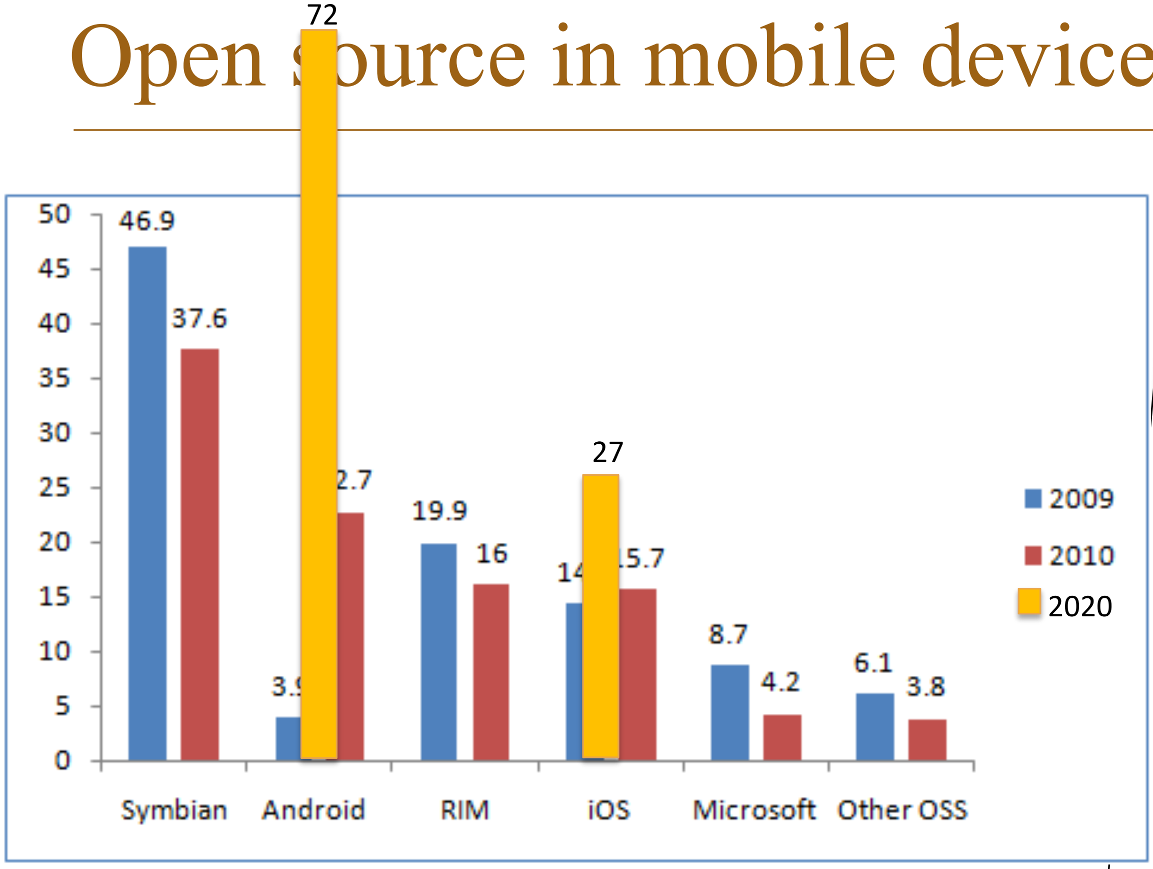
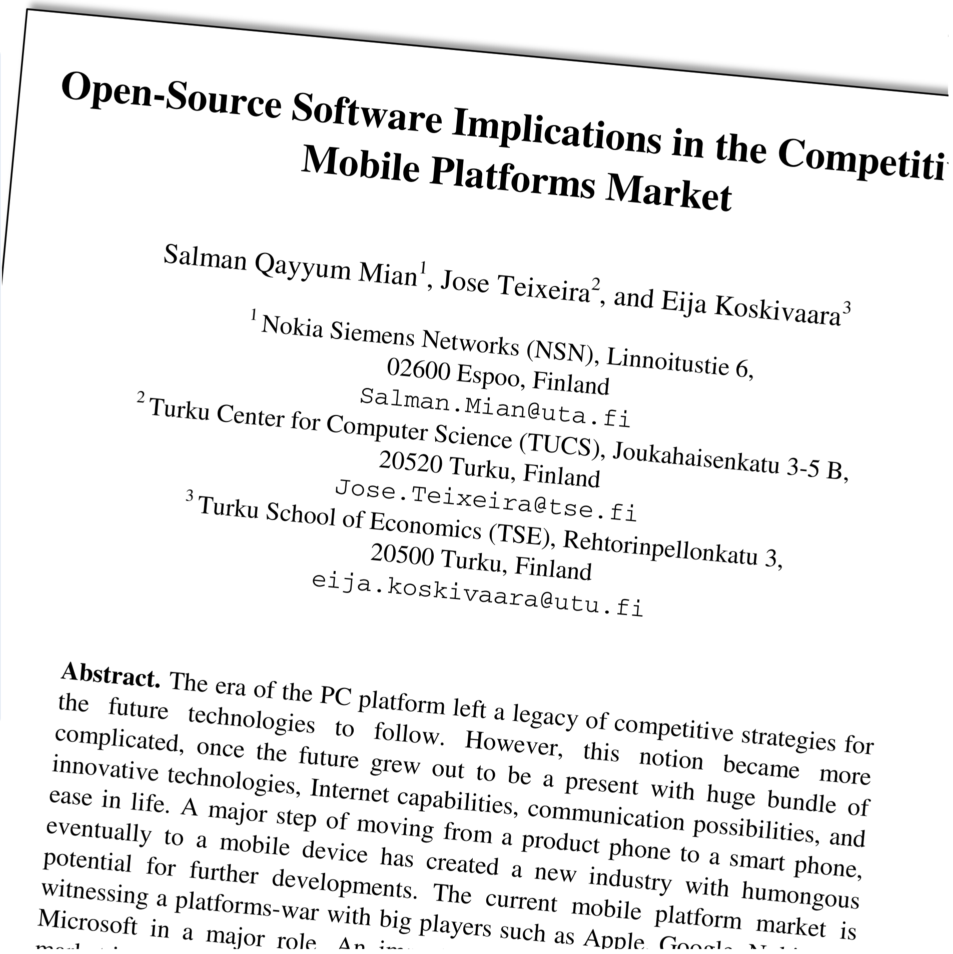


**Fig. 1.** Worldwide smart-phone Market shares (%) by platform in 2009/2010 (Gartner, 2011)



## Open-Source Software Implications in the Competiti Mobile Platforms Market

Salman Qayyum Mian[1], Jose Teixeira[2], and Eija Koskivaara[3]

[1] Nokia Siemens Networks (NSN), Linnoitustie 6,
02600 Espoo, Finland
Salman.Mian@uta.fi

[2] Turku Center for Computer Science (TUCS), Joukahaisenkatu 3-5 B,
20520 Turku, Finland
Jose.Teixeira@tse.fi

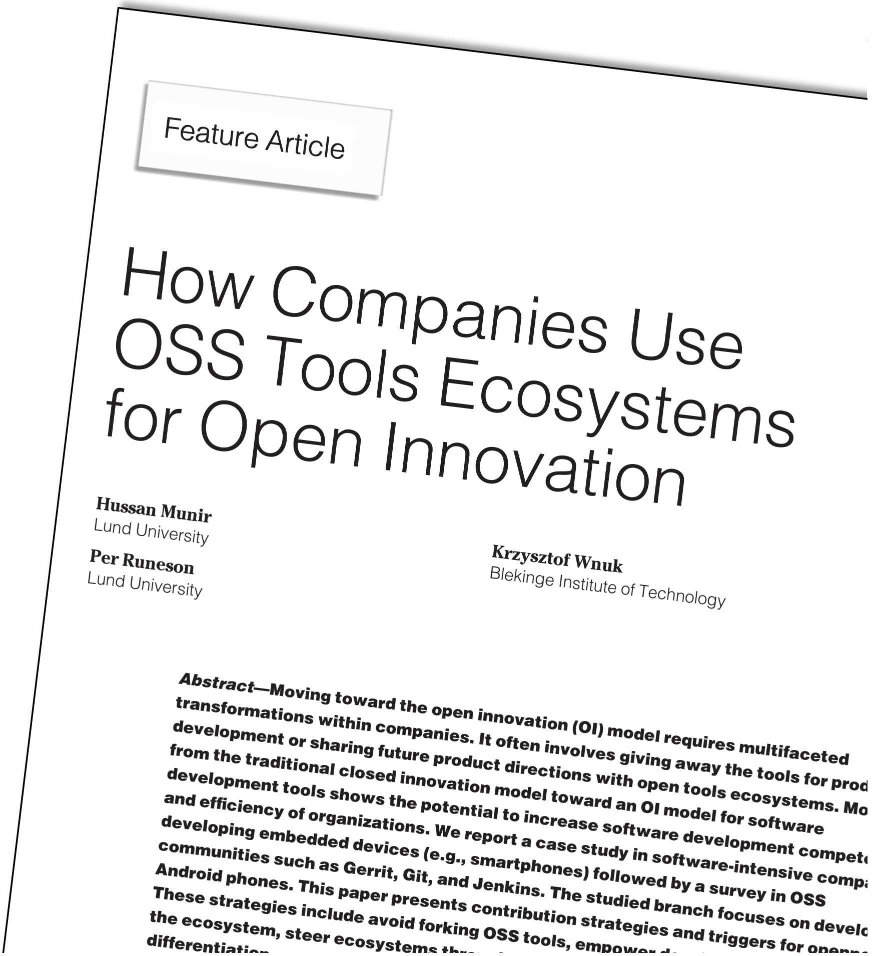
[3] Turku School of Economics (TSE), Rehtorinpellonkatu 3,
20500 Turku, Finland
eija.koskivaara@utu.fi

**Abstract.** The era of the PC platform left a legacy of competitive strategies for the future technologies to follow. However, this notion became more complicated, once the future grew out to be a present with huge bundle of innovative technologies, Internet capabilities, communication possibilities, and ease in life. A major step of moving from a product phone to a smart phone, eventually to a mobile device has created a new industry with humongous potential for further developments. The current mobile platform market is witnessing a platforms-war with big players such as Apple, Google, Nokia and Microsoft in a major role. An im

# Triggers of Openness – why engage?

- Access to skilled workforce
- Faster development speed
- Low license costs and switching costs
- Flexibility in tool usage and adaptations
- Shared cost with the ecosystem
- Governing ecosystem

https://doi.org/10.1109/MITP.2019.2893134



Feature Article

## How Companies Use OSS Tools Ecosystems for Open Innovation

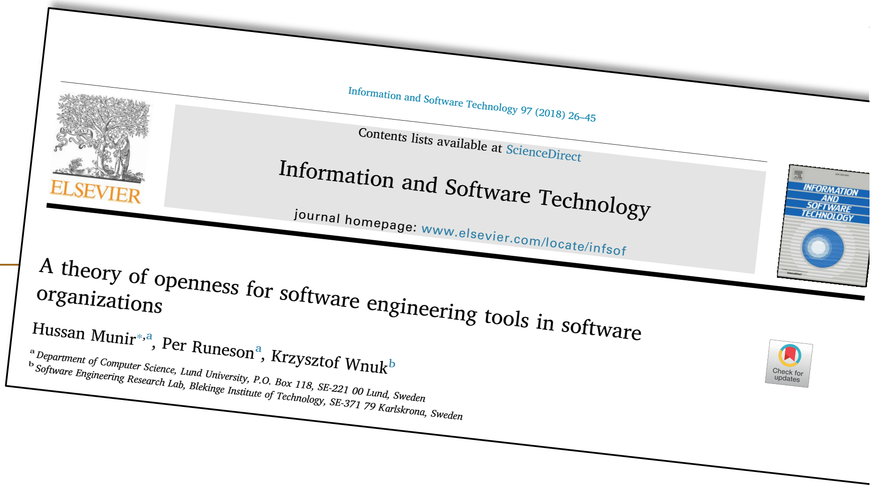
**Hussan Munir**
Lund University

**Per Runeson**
Lund University

**Krzysztof Wnuk**
Blekinge Institute of Technology

***Abstract*—Moving toward the open innovation (OI) model requires multifaceted**
**transformations within companies. It often involves giving away the tools for prod**
**development or sharing future product directions with open tools ecosystems. Mo**
**from the traditional closed innovation model toward an OI model for software**
**development tools shows the potential to increase software development compet**
**and efficiency of organizations. We report a case study in software-intensive compa**
**developing embedded devices (e.g., smartphones) followed by a survey in OSS**
**communities such as Gerrit, Git, and Jenkins. The studied branch focuses on develo**
**Android phones. This paper presents contribution strategies and triggers for openn**
**These strategies include avoid forking OSS tools, empower**
**the ecosystem, steer ecosystems thr**
**differentiation**

# Strategies for open tools



Information and Software Technology 97 (2018) 26–45

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

A theory of openness for software engineering tools in software organizations

Hussan Munir[*,a], Per Runeson[a], Krzysztof Wnuk[b]

[a] Department of Computer Science, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden
[b] Software Engineering Research Lab, Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden

| **Strategy** | Cost saving | Inspirational |
|---|---|---|
| Proactive | **Lucrativeness**<br>*(Think tank)* | **Leaders**<br>(Growth through ecosystems) |
| Reactive | **Laggards**<br>*(Business as usual)* | **Leverage**<br>*(Resource optimization)* |

**Why**

# Open source for data?
# Data Ecosystems!!!

CC BY-ND 2.0 Alick Boych @ Flickr

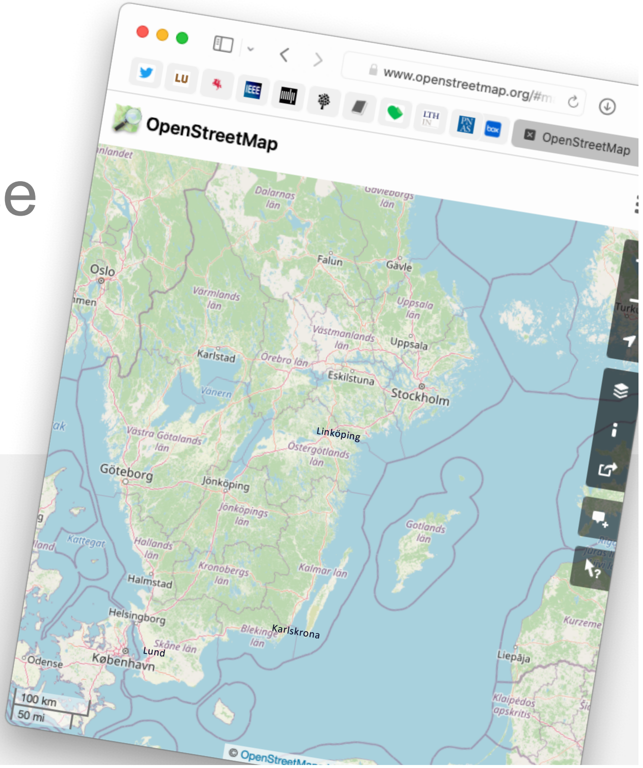
# A Data Ecosystem is...

- a network **community** with a common interest
- supported by a **technological platform**
- to **process data**
  - e.g., find, archive, publish, consume, or reuse
- collaboration on **the data and resources**
  - e.g., software and standards

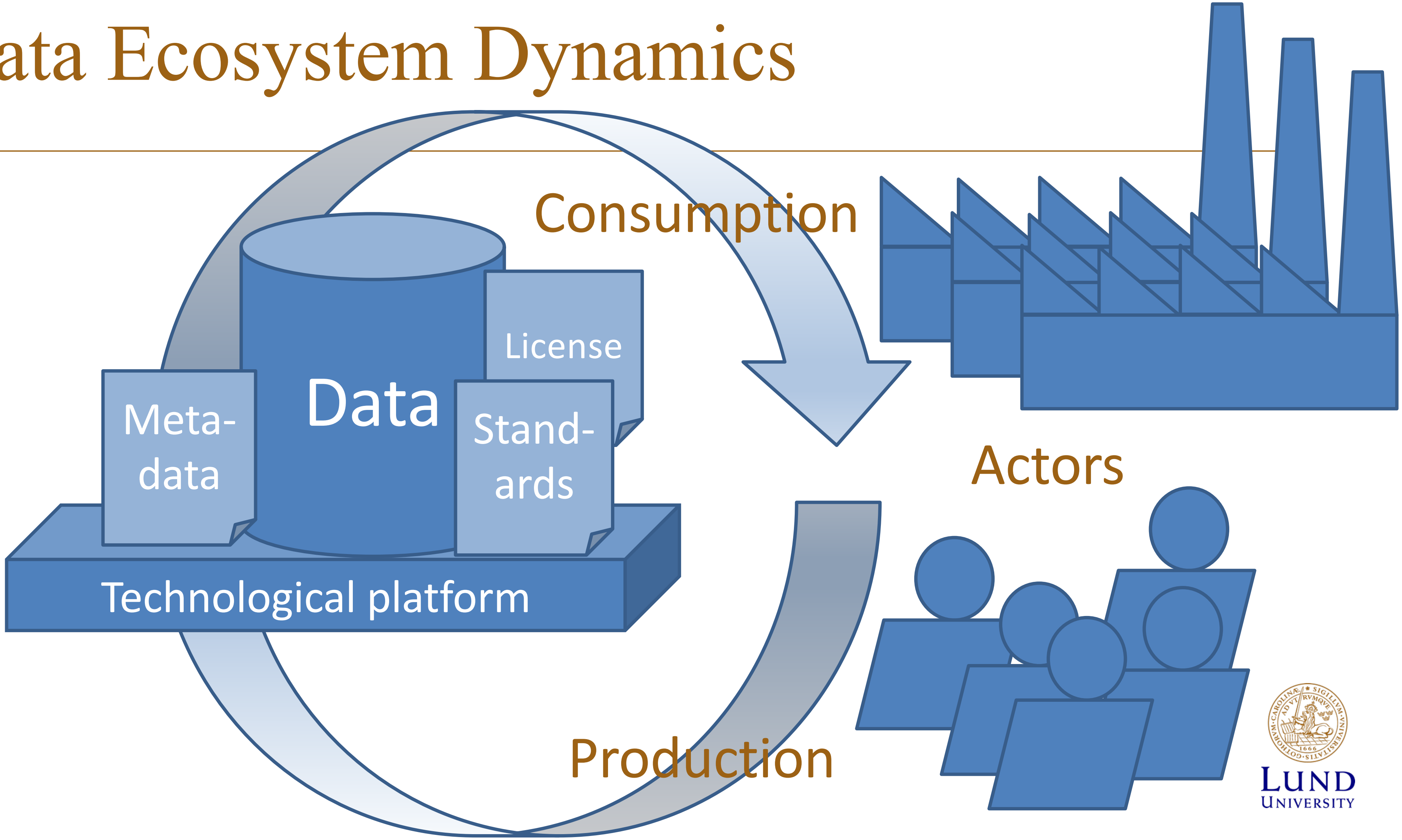
Vi är Trafiklab

Samtrafiken TRAFIKVERKET SL ÖstgötaTrafiken UL

västtrafik Skånetrafiken RISE

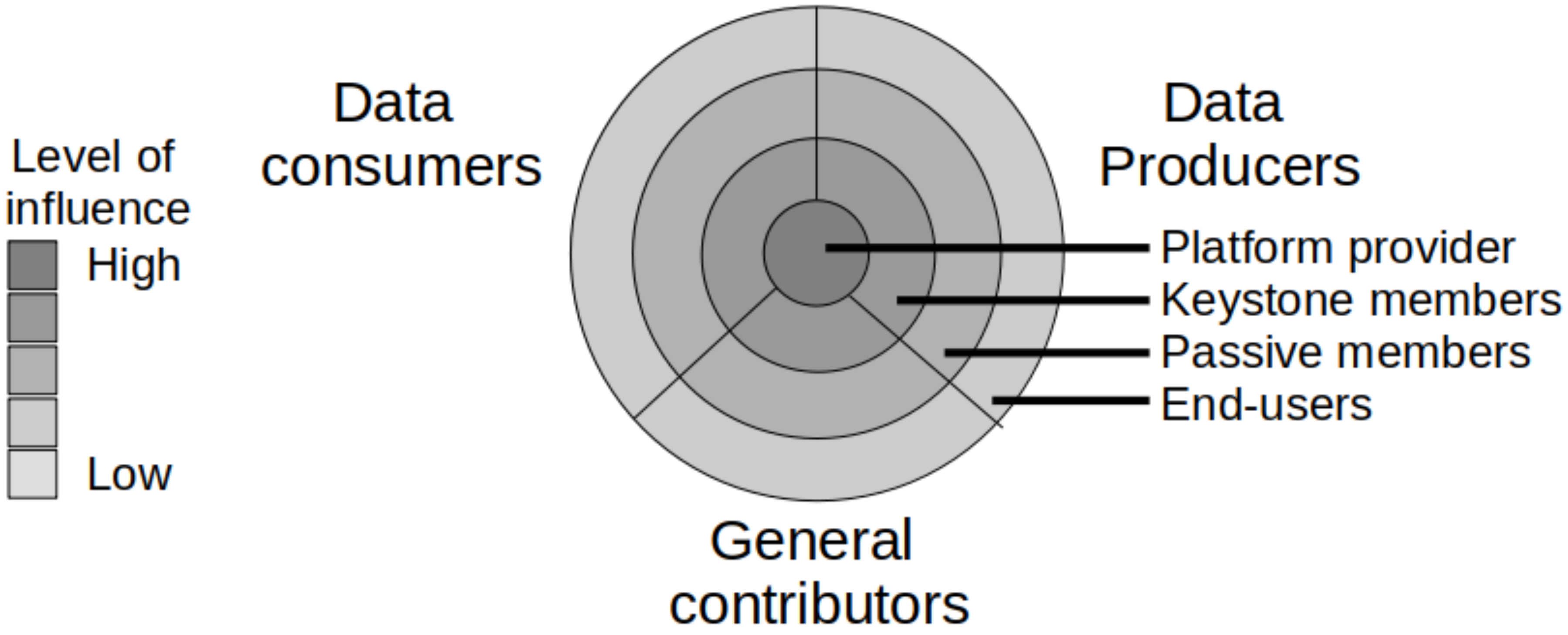


# Data Ecosystem Dynamics



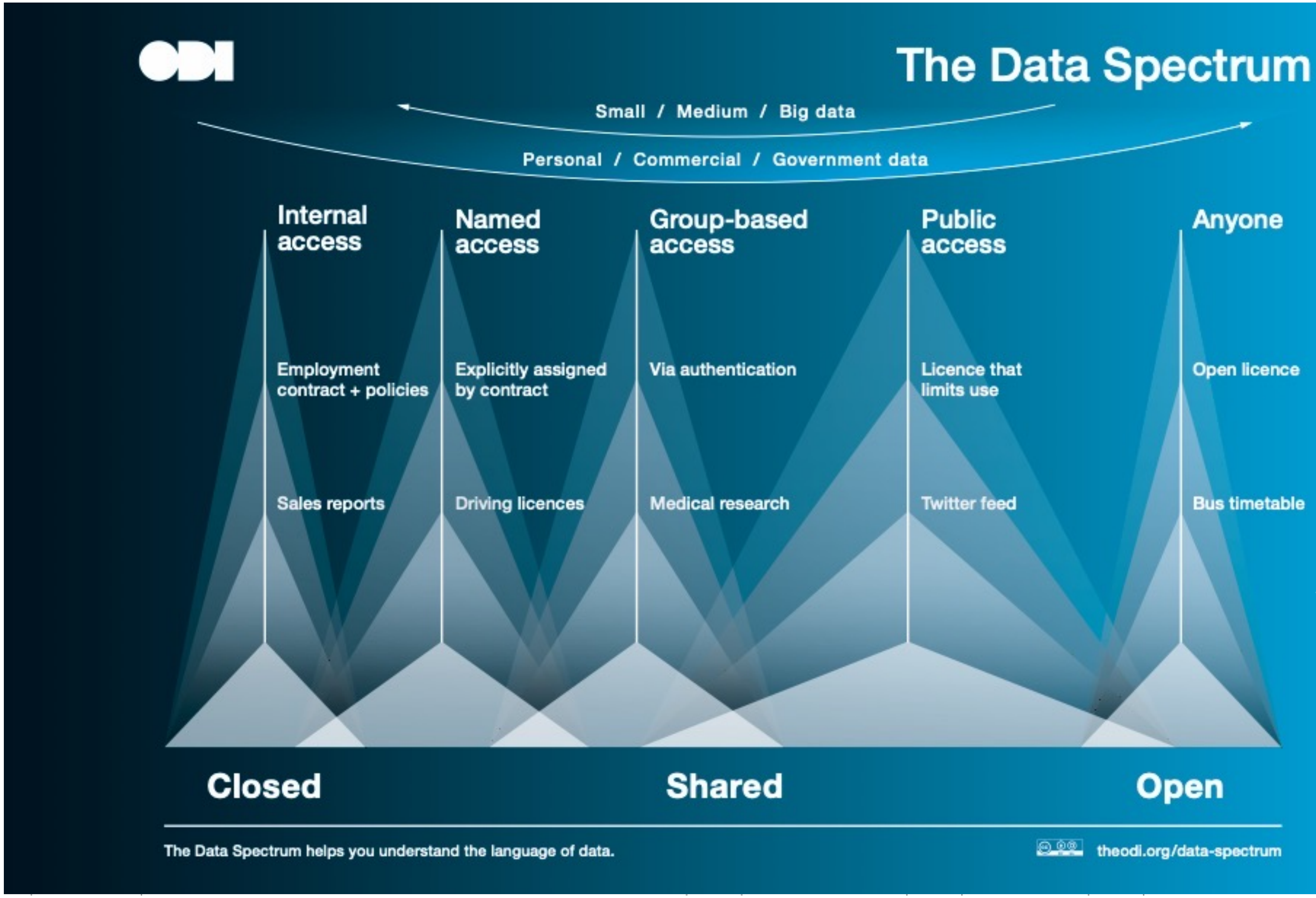
LUND UNIVERSITY

# Data Ecosystem Roles



https://doi.org/10.1007/978-3-030-57599-1%5C_22

# How open is open?



ODI

## The Data Spectrum

Small / Medium / Big data

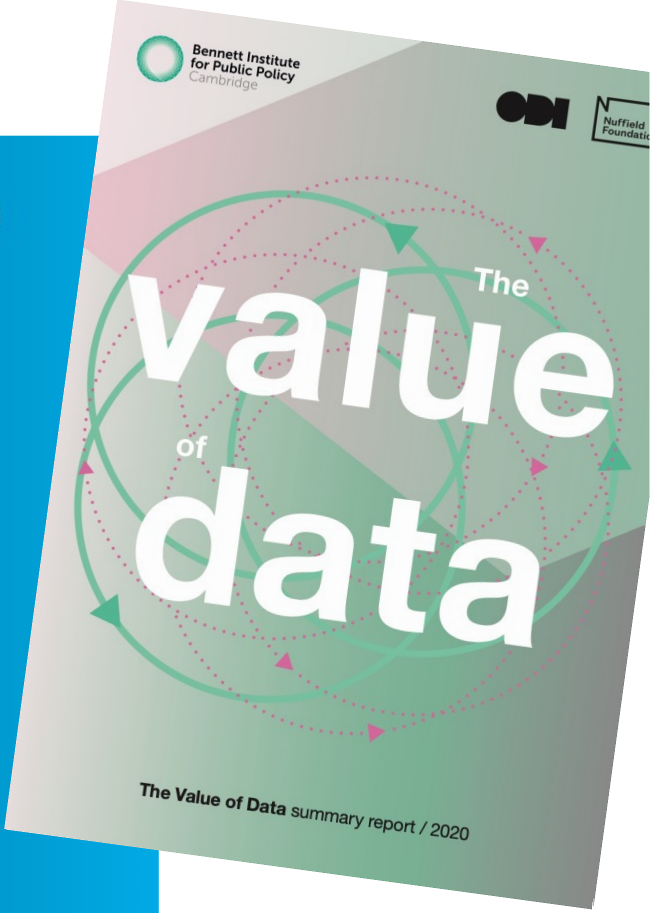
Personal / Commercial / Government data

| Internal access | Named access | Group-based access | Public access | Anyone |
|---|---|---|---|---|
| Employment contract + policies | Explicitly assigned by contract | Via authentication | Licence that limits use | Open licence |
| Sales reports | Driving licences | Medical research | Twitter feed | Bus timetable |

Closed — Shared — Open

The Data Spectrum helps you understand the language of data.

theodi.org/data-spectrum



Bennett Institute for Public Policy Cambridge

ODI

Nuffield Foundatio

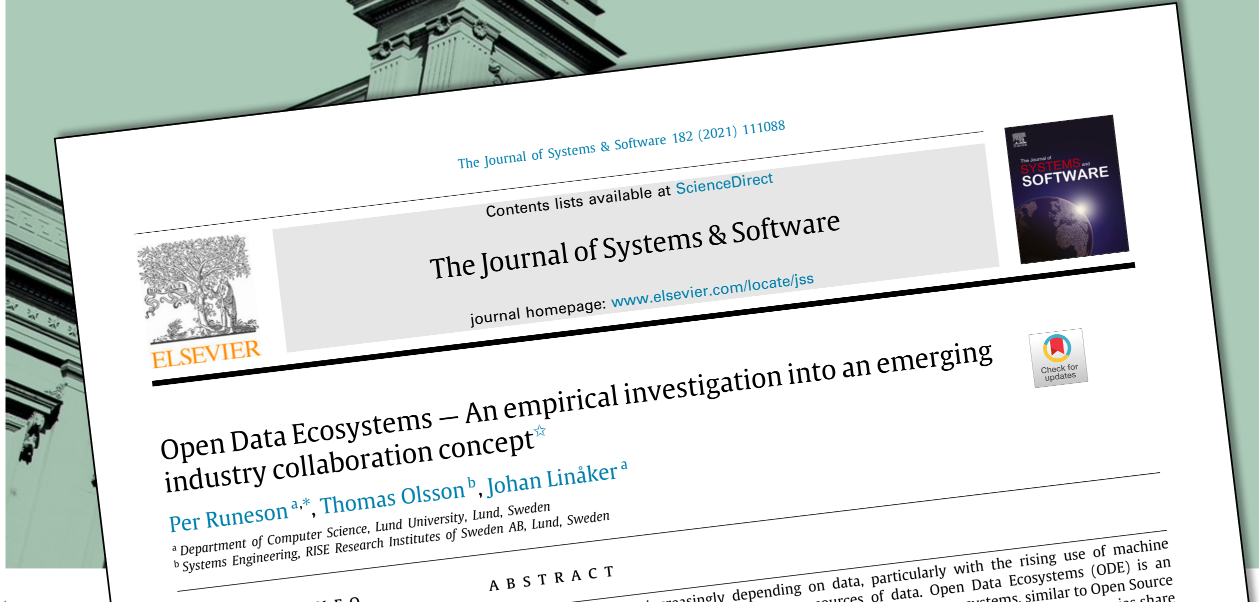
The value of data

The Value of Data summary report / 2020

LUND UNIVERSITY

# Essential concepts of Open Data Ecosystems

PER RUNESON, THOMAS OLSSON, JOHAN LINÅKER



The Journal of Systems & Software 182 (2021) 111088

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

Open Data Ecosystems — An empirical investigation into an emerging industry collaboration concept☆

Per Runeson [a,*], Thomas Olsson [b], Johan Linåker [a]

[a] Department of Computer Science, Lund University, Lund, Sweden
[b] Systems Engineering, RISE Research Institutes of Sweden AB, Lund, Sweden

A B S T R A C T

depending on data, particularly with the rising use of machine
of data. Open Data Ecosystems (ODE) is an
similar to Open Source

# Emerging data ecosystems

**JobTech**

- Labor market
- Job ads
- Public-driven
- Organization-centric

**ESS-CSDL**

- Industry 4.0
- Alarm data
- Business-driven
- Organization-centric

**RoDL**

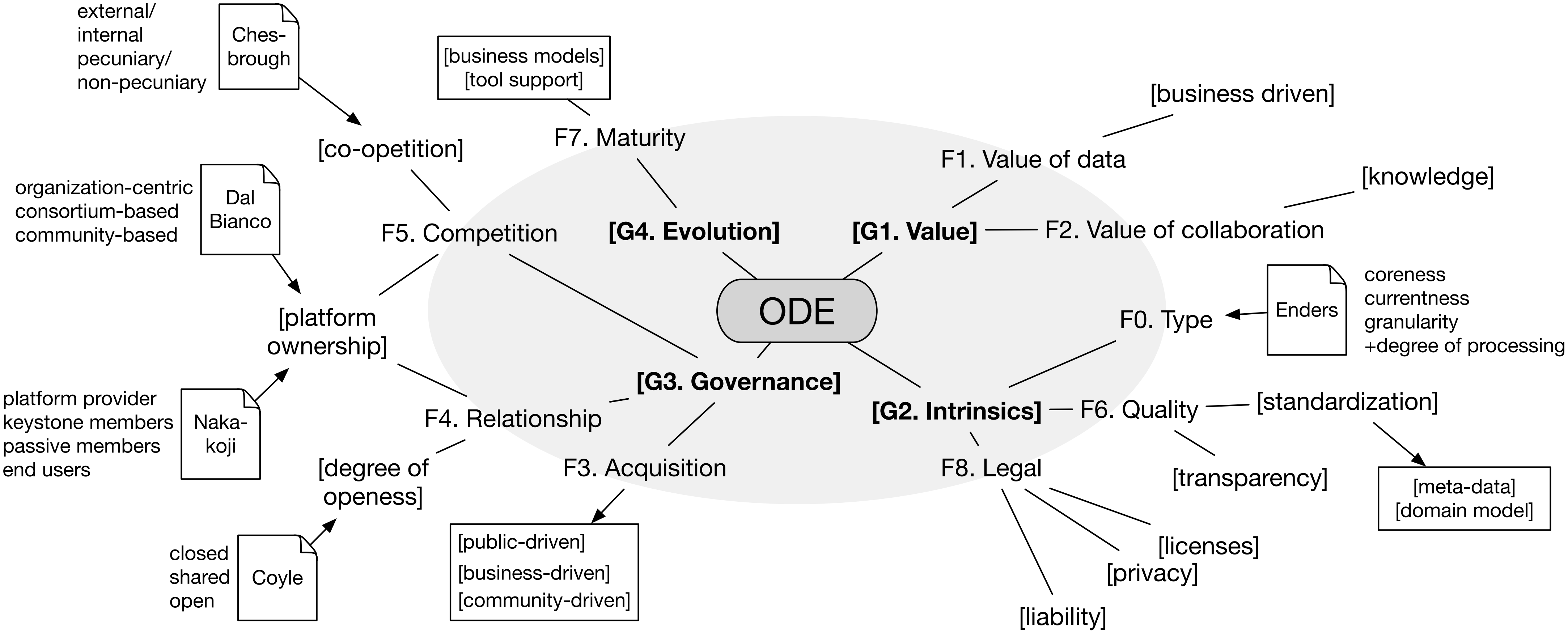
- Automotive
- Traffic video
- Business-driven
- Consortium-based

JobTech Development

ess EUROPEAN SPALLATION SOURCE

101 0010
110 1111
100 0100
100 1100
RoDL

LUND UNIVERSITY

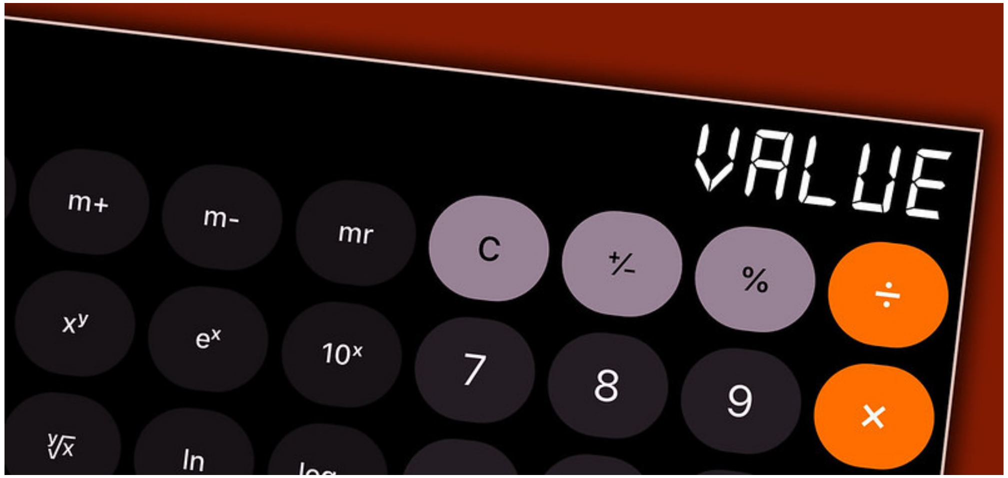


Open Data Ecosystems – an empirical investigation into an emerging industry collaboration concept

LUND UNIVERSITY

# Value

The value of data (F1) and the value of collaboration around the data (F2) are two sides of the same coin. One or the other may be the primary value, but they are highly intertwined.



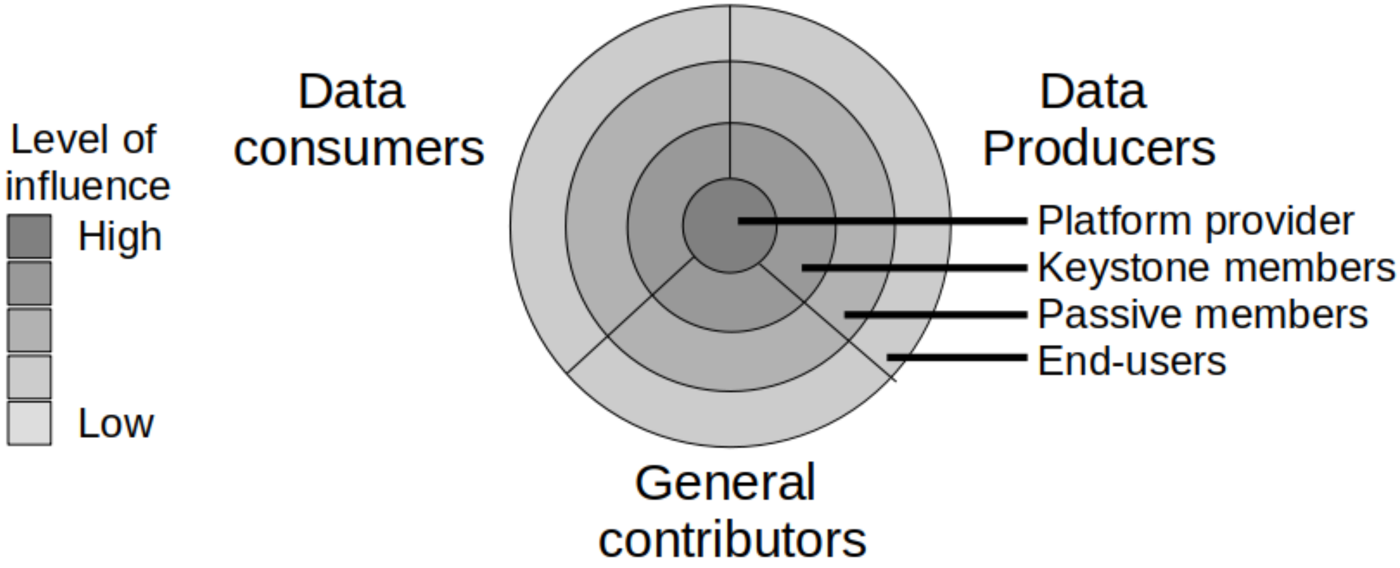
CC-BY 2.0 Mike Lawrence @Flickr

# Intrinsics

Intrinsics,
or internal characteristics of data

- data type (F0)
  - coreness
  - currentness
  - granularity
  - degree of processing
- data quality (F6)
  - correctness
  - provenance
  - meta-data
- legal aspects (F8) is tightly connected to data, although they also connect to governance of the ODE.
  - licenses
  - privacy
  - liability

# Governance



There is a need for an independent platform provider to ensure trust

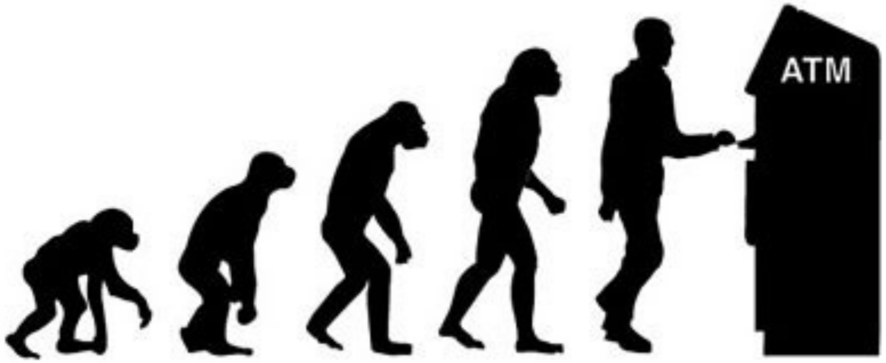
Initiation may be public-driven, business-driven, or community-driven

# Evolution

The concept of and strategies for open data ecosystems are still in their infancy

Need for knowledge:

- how to integrate ODEs into an organization’s business model
- tools to support ODEs and enable data sharing should be developed and standardized



(CC BY-NC 2.0) ThomasThomas@Flickr

# Findings for data ecosystems

**Value**

Focus on business value in the data or collaboration

**Intrinsics**

Data coreness, currentness and granularity

Standardize format and legal framework

**Governance**

Level of openness and platform ownership

Relationship and competition must co-exist

Data acquisition incentives

**Evolution**

Advance business models and tool support

CC BY-SA 2.0 Jocelyn Kinghorn @ Flickr

# Recommendations for public platform providers

JOHAN LINÅKER, PER RUNESON

JeDEM Issue 13 (1): 1-30, 2021
ISSN 2075-9517
http://www.jedem.org
Date of submission: 29.01.2021
Date of acceptance: 20.04.2021
Article DOI: 10.29379/jedem.v13i1.634

## How to Enable Collaboration in Open Government Data Ecosystems: A Public Platform Provider's Perspective

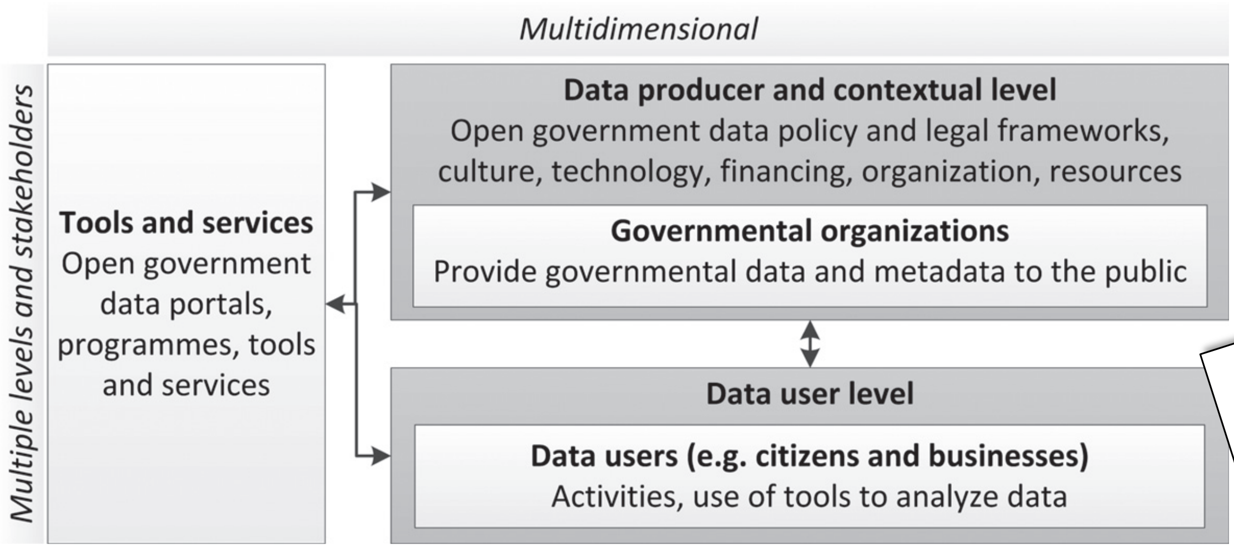
Johan Linåker and Per Runeson

Dept. of Computer Science, Lund University, Lund (SWE), johan.linaker@cs.lth.se, per.runeson@cs.lth.se

Open Government Data (OGD) is an important driver for open innovation among public
arch highlights a need for improved feedback loops, collaboration,
study, we explore how public platform
both in terms

# Open Governmental Data

Purpose: 1) Governance transparency, 2) Business innovation



*Multidimensional*

*Multiple levels and stakeholders*

**Tools and services**
Open government data portals, programmes, tools and services

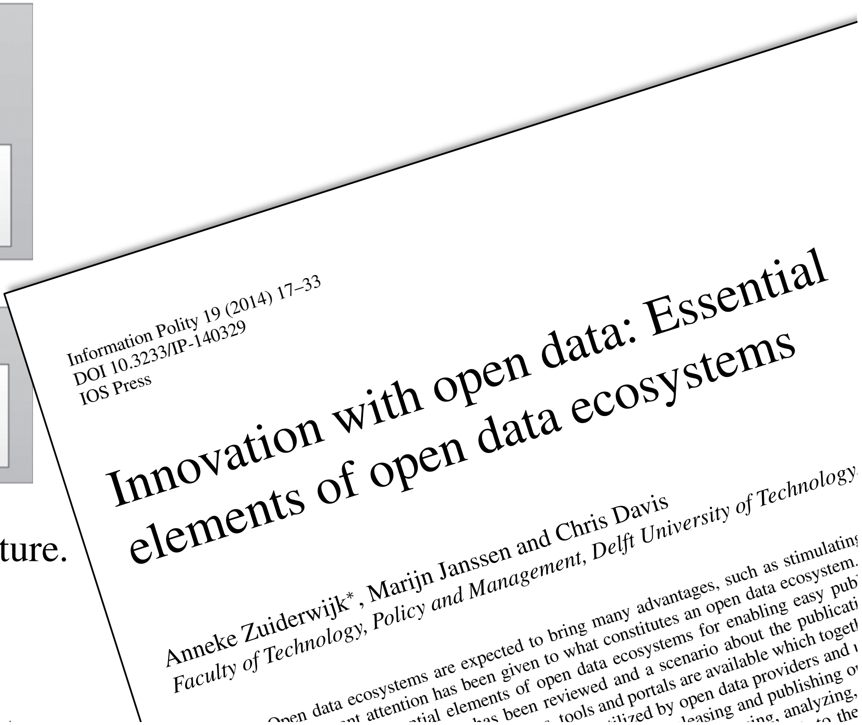
**Data producer and contextual level**
Open government data policy and legal frameworks, culture, technology, financing, organization, resources

**Governmental organizations**
Provide governmental data and metadata to the public

**Data user level**

**Data users (e.g. citizens and businesses)**
Activities, use of tools to analyze data

Fig. 1. Elements of an open government data ecosystem derived from the literature.



Information Polity 19 (2014) 17–33
DOI 10.3233/IP-140329
IOS Press

Innovation with open data: Essential elements of open data ecosystems

Anneke Zuiderwijk*, Marijn Janssen and Chris Davis
*Faculty of Technology, Policy and Management, Delft University of Technology*

Open data ecosystems are expected to bring many advantages, such as stimulating
nt attention has been given to what constitutes an open data ecosystem.
ntial elements of open data ecosystems for enabling easy pub
has been reviewed and a scenario about the publicati
tools and portals are available which toget
tilized by open data providers and u
leasing and publishing o
ing, analyzing,
to the

# Open Government Data ecosystems

### JobTech

- Labor market
- Job ads
- Public-driven
- Organization-centric

### Trafiklab

- Public transport
- Schedule, traffic
- Public-driven
- Consortium-based

### HSL DevCom

- Public transport
- Schedule, traffic
- Public-driven
- Organization-centric

### City of Chicago

- City governance
- All kinds of city
- Public-driven
- Organization-centric

JobTech Development

TRAFIKLAB, A BRAND OF SAMTRAFIKEN

Samtrafiken

HSL HRT

CHICAGO

# Recommendations for the public platform provider

Build community and trust

Maintain a clear vision

Be active and multi-functional

Build open communication

Develop open source software

Share data, other than your own

Adopt and promote open standards

CC BY-SA 2.0 Jocelyn Kinghorn @ Flickr

LUND
UNIVERSITY

# Open data ecosystems – wishful thinking or successful business?

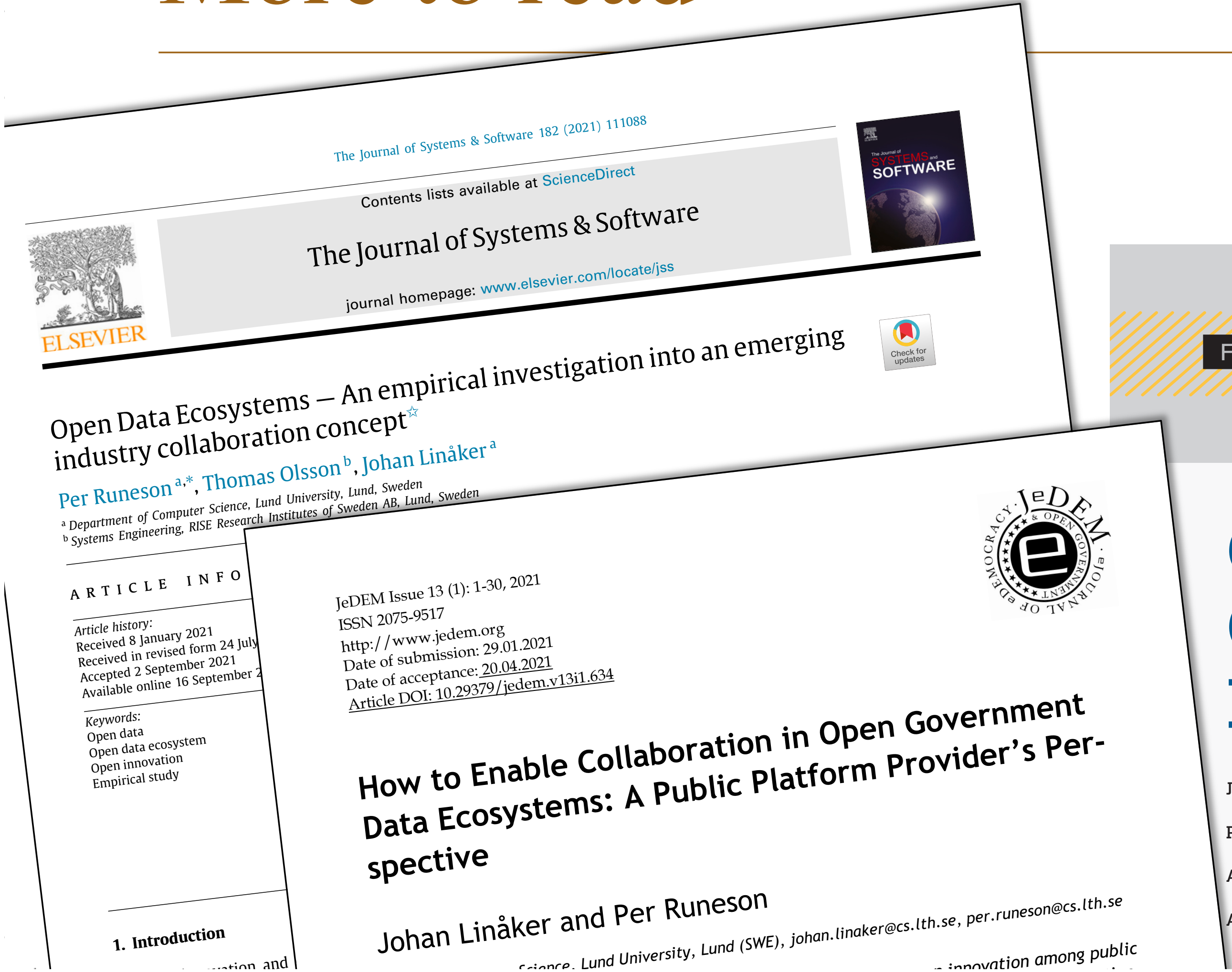
# Maybe data isn't the new oil?

It might be the new, renewable bio-energy
but we have to make it together

CC BY-ND 2.0 Alick Boych @ Flickr

# More to read

https://doi.org/10.1016/j.jss.2021.111088
https://doi.org/10.29379/jedem.v13i1.634



The Journal of Systems & Software 182 (2021) 111088

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

ELSEVIER

Open Data Ecosystems — An empirical investigation into an emerging industry collaboration concept☆

Per Runeson [a],*, Thomas Olsson [b], Johan Linåker [a]

[a] Department of Computer Science, Lund University, Lund, Sweden
[b] Systems Engineering, RISE Research Institutes of Sweden AB, Lund, Sweden

ARTICLE INFO

Article history:
Received 8 January 2021
Received in revised form 24 July
Accepted 2 September 2021
Available online 16 September 2

Keywords:
Open data
Open data ecosystem
Open innovation
Empirical study

1. Introduction

JeDEM Issue 13 (1): 1-30, 2021
ISSN 2075-9517
http://www.jedem.org
Date of submission: 29.01.2021
Date of acceptance: 20.04.2021
Article DOI: 10.29379/jedem.v13i1.634

How to Enable Collaboration in Open Government Data Ecosystems: A Public Platform Provider's Perspective

Johan Linåker and Per Runeson

Science, Lund University, Lund (SWE), johan.linaker@cs.lth.se, per.runeson@cs.lth.se

innovation among public

FOCUS: GUEST EDITORS' INTRODUCTION

## Collaborative Aspects of Open Data in Software Engineering

**Johan Linåker**, RISE Research Institutes of Sweden

**Per Runeson**, Lund University

**Anneke Zuiderwijk**, Delft University of Technology

**Amanda Brock**, OpenUK

Upcoming IEEE Software January/February 2022

# More to come:
# B2B Data Sharing for Industry 4.0 Machine Learning

Kowalkowski

Ahmed

Prof. **Per Runeson**, PhD Student **Konstantin Malysh**, Software Engineering, LU
Prof. **Christian Kowalkowski**, PhD Student **Tanvir Ahmed**, Industrial Marketing, LiU

## Business models (LiU)

Two disruptive and interrelated transformations:

1) **digitalization** changes sociotechnical systems,
2) **servitization** entails the shift from selling products to 'product-as-a-service' business models

## Collaboration tools (LU)

Git, Jenkins and Gerrit, provide a low-threshold entry  o open source software (OSS). Data ecosystems need "an underpinning technological platform".

Runeson

Malysh

ELLIIT

Excellence Center at Linköping – Lund in Information Technology

LUND UNIVERSITY

# WE ARE OPEN

per.runeson@cs.lth.se
http://www.lth.se/digitalth