PARCOS

Participatory Communication of Science

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 872500.

Open Research and Data Analysis: European Project Context

Antti Knutas
Natasha Tylosky
Tanvir Hasan

LUT University

Today’s tutorial content

Discussion of open data requirements from a research project perspective, briefly recap open science data, and present a sample workflow that can assist you in opening up your both analysis process and data.

- Discussion of open data and presentation of workflow (now, Knutas)
- Demo of one workflow (next, Knutas)
- Tutorial of getting started (Tanvir Hasan)
- Testing of tools and Q&A
- Coffee break
- Demo of web visualization and getting your web content archived openly (Natasha Tylosky)
- Mapping of challenges, discussion of participants’ research processes and conclusion (everyone)

Pt. A: Open Science Definitions and Advantages

What is Open Science?

As defined by EC (https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science_en)

To improve quality, efficiency and responsiveness of research…

- “When researchers share knowledge and data as early as possible in the research process with all relevant actors it helps diffuse the latest knowledge.”
- “And when partners from across academia, industry, public authorities and citizen groups are invited to participate in the research and innovation process, creativity and trust in science increases.”
- Also: Transparency and reproducibility

Caveat: “The effective linking of open science practices to innovation and business models requires careful consideration of issues such as Intellectual Property Rights (IPR), licensing agreements, interoperability and reuse of data.”

What is Open Science?

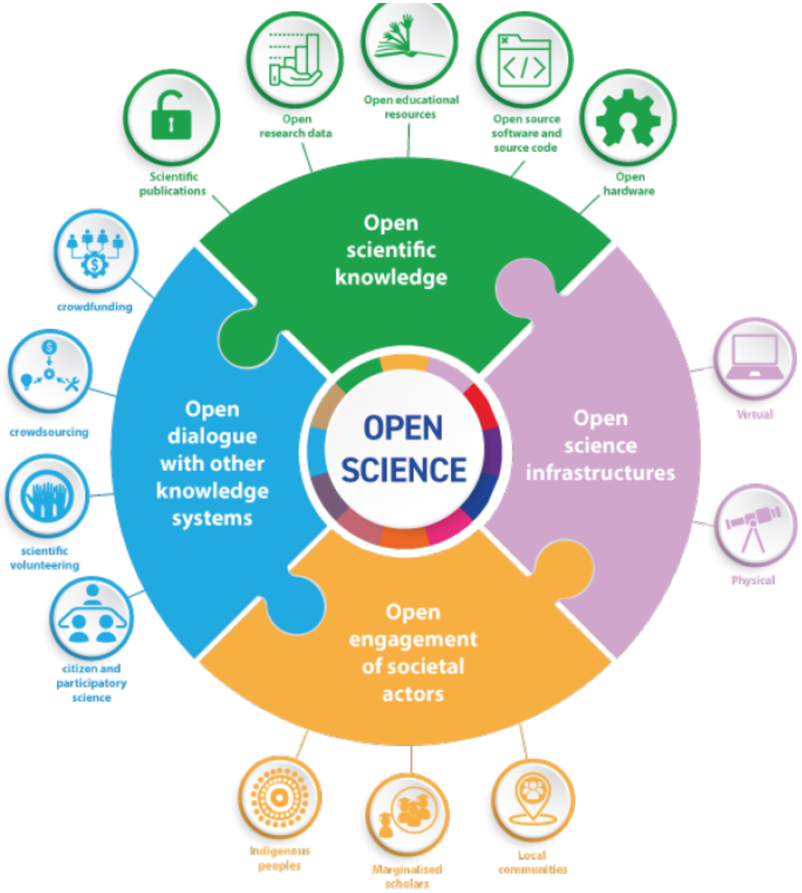
As defined by UNESCO (https://en.unesco.org/science-sustainable-future/open-science/recommendation)

To ensure science truly benefits the people and the planet, **Open Science** is a movement to make science more open, accessible, efficient, democratic and transparent.

- Open Access
- Open Data
- Open to Society

Pillars

As presented by UNESCO


Scientific publications
Open research data
Open educational resources
Open source software and source code
Open hardware
Open scientific knowledge
crowdfunding
crowdsourcing
scientific volunteering
citizen and participatory science
Open dialogue with other knowledge systems
OPEN SCIENCE
Open science infrastructures
Virtual
Physical
Open engagement of societal actors
Indigenous peoples
Marginalised scholars
Local communities

Open Scientific Knowledge: scientific publications, research data, software, source code and hardware available in the public domain or under the copyright that has been released under an open license

Open Science infrastructures: scientific equipment or sets of instruments, knowledge-based resources such as collections, repositories, archives and scientific data, open computational and digital infrastructures, needed to support Open Science and serve the needs of different communities

Open engagement of societal actors: citizen and participatory science and other extended collaboration between scientists and societal actors beyond the scientific community, opening up practices and tools that are part of the research cycle and by making the scientific process more inclusive and accessible to the broader inquiring society

Open dialogue with other knowledge systems: recognition of complementarities between diverse epistemologies, including indigenous knowledge systems

Open science is not just open data – that’s the minimum level

…but the open science data publishing process is the topic of this workshop.

Kindly catch me during a coffee break, lunch or dinner to hear more, if you’re interested!

Why open science?

List of diverse reasons

Improving efficiency, access, and responsivity

- Transparency and replicability of science
- Reusability and knowledge transfer
- Accessibility to resources (public research, public access?)
- Productivity ("open innovation")
- Building trust with people (engaging science)

…or your funder made you do it!

Downsides?

- Cost in time and effort
- Copyright and other barriers => (software) sales related to long-term sustainability of research artefacts
- Fear of someone “taking” your results?
- GDPR and ethical aspects in quantitative data
- Difficulties in anonymizing qualitative data

Let’s discuss these and try to address some at the end

Pt. B: Open data requirements found in research projects

Antti Knutas

Some considerations and standards

- Data quality and access (“FAIR” principles)
- Licensing
- How to document your open science? (Data Management Plan)
- Process & tools?

Open Data and FAIR principles

Findable, Accessible, Interoperable, Reusable (https://www.go-fair.org/fair-principles/)

- Findable – metadata, discovery

Example: Zenodo metadata (not just as a ZIP / PDF in website)

- Accessible – openly available

Example: Web protocols, APIs

- Interoperable – data exchange, standards, vocabularies

Example: Standard data formats, description of fields

- Reusable – licensed to permit reuse and rich documentation

Example: Creative Commons licensing, documenting the data

Licensing

Open science: Also open to reuse

- Licenses: A simple, standardised way to allow other people to share, modify and use the research outcomes
- Often allow reuse and modifying (free and open source), but require attribution
- It depends on the license type whether derivate works need to use same license or can be closed (largest controversy in F/OSS discussion)
- How to select a license? Creativecommons.org has a guide and a wizard
- (for software artefacts, MIT and GPLv3 are common)

Disclaimer: F/OSS as a topic raises a lot of passions. In this presentation, we approach it from an utilitarian fashion.

Our case: ORDP in a Horizon 2020, EU funded project

Open Research Data Pilot

The conditions we adhere to, are:

- Develop (and keep up-to-date) a Data Management Plan (DMP).
- Deposit your data in a research data repository.
- Ensure third parties can freely access, mine, exploit, reproduce and disseminate your data.
- Provide related information and identify (or provide) the tools needed to use the raw data to validate your research.

Pilot applies to:

- The data (and metadata) needed to validate results in scientific publications.
- Other curated and/or raw data (and metadata) that you specify in the DMP.

Some further personal motivation

Our funder has kindly requested us to both publish openly (open access publications)

Furthermore, our university counts publications ONLY

- The data (and metadata) needed to validate results in scientific publications.
- Other curated and/or raw data (and metadata) that you specify in the DMP.

…and yours truly is committed to open & engaging science as a principle, but the previous two listed reasons listed above are great motivators.

Data Management Plan

EC on FAIR data management -
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Questions that we have addressed, briefly: What standards / licenses will be applied?

  - (Creative Commons 4.0 BY, MIT and CDLA)

- How data will be exploited and/or shared/made accessible for verification and reuse?
  - ParCos platform as a key portal to open materials, including GitHub, Zenodo, and different media platforms
- How data will be curated and preserved?
  - ParCos platform during project, Zenodo as long-term storage

https://parcos-project.eu/wp-content/uploads/2021/03/D1.1-Data-Management-Plan.pdf

Pt. C: Workflow

Antti Knutas

Overview

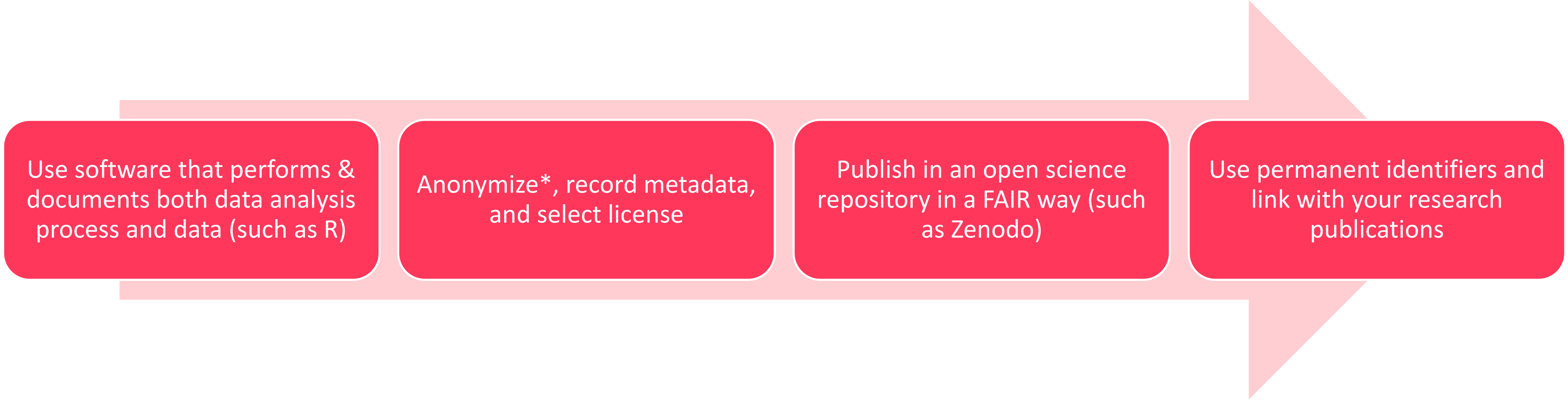
Reviewing the process and then demonstrating it

Sample workflow with…

- Helsinki Region Infoshare open data
- RStudio environment
- Rmarkdown documents
- Zenodo open access repository

Open data workflow

How to create a unified process without *too much* undue work?



Use software that performs & documents both data analysis process and data (such as R)

Anonymize*, record metadata, and select license

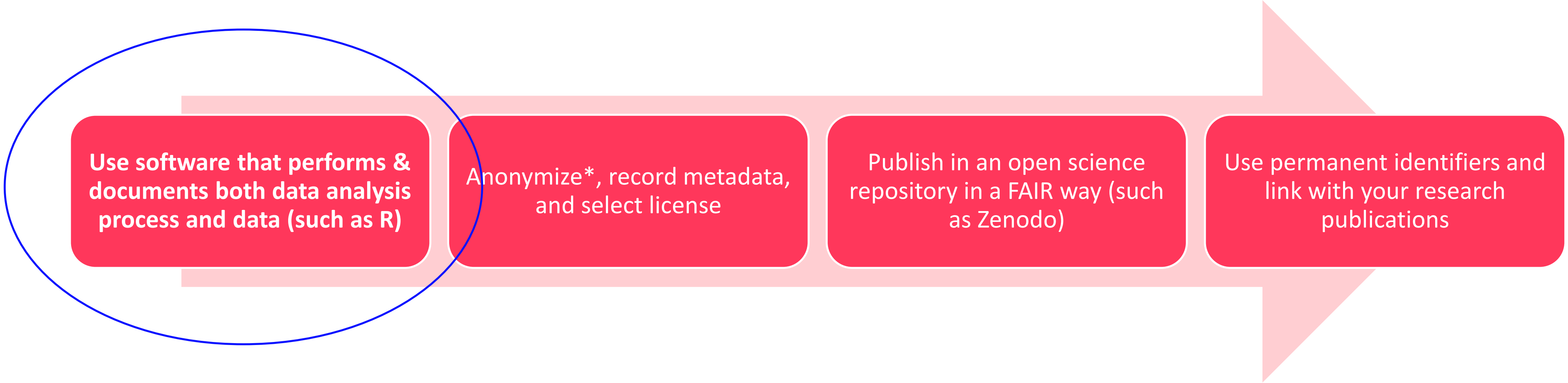
Publish in an open science repository in a FAIR way (such as Zenodo)

Use permanent identifiers and link with your research publications

* As open as possible, as closed as necessary (GDPR, commercial interests, vulnerable groups etc.)

Open data workflow: Step 1

Use software that…



Use software that performs & documents both data analysis process and data (such as R)

Anonymize*, record metadata, and select license

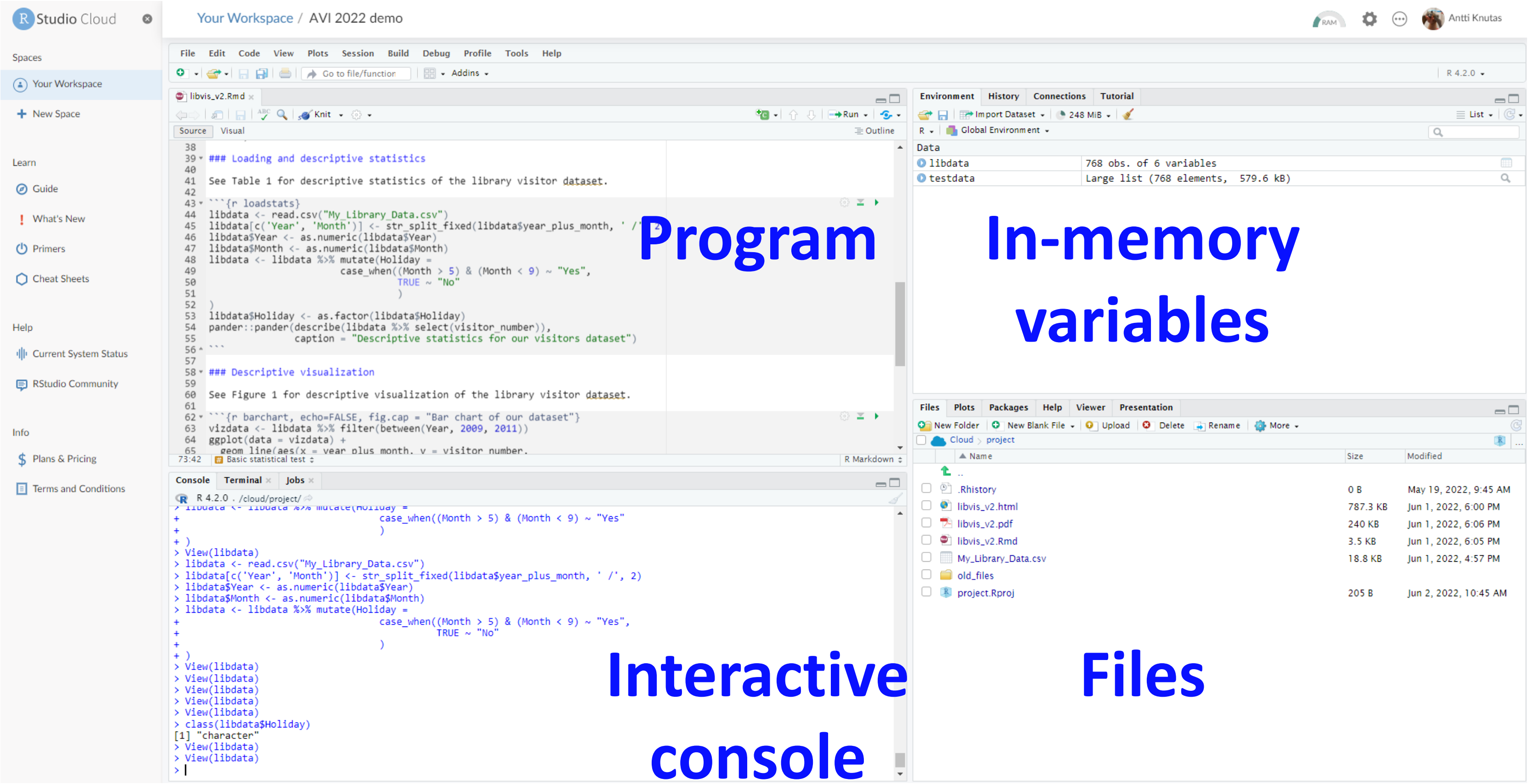
Publish in an open science repository in a FAIR way (such as Zenodo)

Use permanent identifiers and link with your research publications

* As open as possible, as closed as necessary (GDPR, commercial interests, vulnerable groups etc.)

R and RStudio

Statistics programming environment



~~~
### Loading and descriptive statistics

See Table 1 for descriptive statistics of the library visitor dataset.

```{r loadstats}
libdata <- read.csv("My_Library_Data.csv")
libdata[c('Year', 'Month')] <- str_split_fixed(libdata$year_plus_month, ' /', 2
libdata$Year <- as.numeric(libdata$Year)
libdata$Month <- as.numeric(libdata$Month)
libdata <- libdata %>% mutate(Holiday =
                        case_when((Month > 5) & (Month < 9) ~ "Yes",
                                  TRUE ~ "No"
                                  )
)
libdata$Holiday <- as.factor(libdata$Holiday)
pander::pander(describe(libdata %>% select(visitor_number)),
               caption = "Descriptive statistics for our visitors dataset")
~~~

Descriptive visualization

See Figure 1 for descriptive visualization of the library visitor dataset.

```{r barchart, echo=FALSE, fig.cap = "Bar chart of our dataset"}
vizdata <- libdata %>% filter(between(Year, 2009, 2011))
ggplot(data = vizdata) +
  geom_line(aes(x = year_plus_month, y = visitor_number,
```

```
> libdata <- libdata %>% mutate(Holiday =
+                                   case_when((Month > 5) & (Month < 9) ~ "Yes"
+                                   )
+ )
> View(libdata)
> libdata <- read.csv("My_Library_Data.csv")
> libdata[c('Year', 'Month')] <- str_split_fixed(libdata$year_plus_month, ' /', 2)
> libdata$Year <- as.numeric(libdata$Year)
> libdata$Month <- as.numeric(libdata$Month)
> libdata <- libdata %>% mutate(Holiday =
+                                   case_when((Month > 5) & (Month < 9) ~ "Yes",
+                                             TRUE ~ "No"
+                                   )
+ )
> View(libdata)
> View(libdata)
> View(libdata)
> View(libdata)
> View(libdata)
> class(libdata$Holiday)
[1] "character"
> View(libdata)
> View(libdata)
> 
```

Data

libdata	768 obs. of 6 variables
testdata	Large list (768 elements, 579.6 kB)

Name	Size	Modified
..		
.Rhistory	0 B	May 19, 2022, 9:45 AM
libvis_v2.html	787.3 KB	Jun 1, 2022, 6:00 PM
libvis_v2.pdf	240 KB	Jun 1, 2022, 6:06 PM
libvis_v2.Rmd	3.5 KB	Jun 1, 2022, 6:05 PM
My_Library_Data.csv	18.8 KB	Jun 1, 2022, 4:57 PM
old_files		
project.Rproj	205 B	Jun 2, 2022, 10:45 AM

Program

In-memory variables

Interactive console

Files

Rmarkdown projects

Combine data description, data analysis commands, and output

- Intertwines commands from...
  - R (language for statistical programming)
  - Markdown, a markup language
  - Data visualizers, such as ggplot

Why => combines documentation, analysis commands *and* the output

=> document outputs as PDF/html/docx

Rmarkdown demo

Let’s have a look at our project

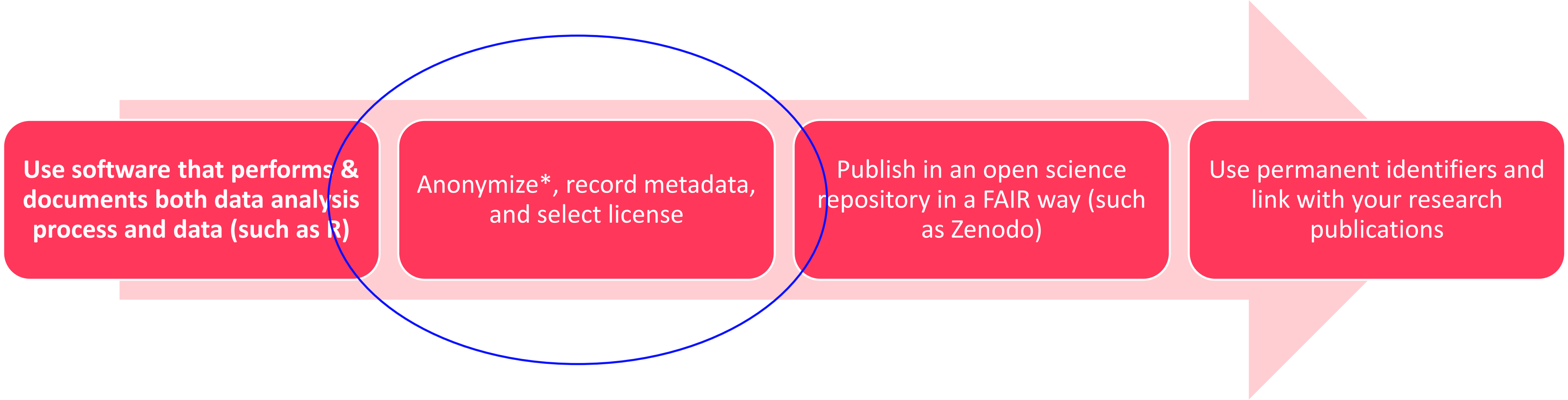
(link in workshop page)

…you could technically do the same in Jupyter & Python, JASP, PSPP (if you document commands & output)

…or even SPSS or Stata (but the execution environment wouldn’t be open)

Open data workflow: Step 2

Select license and anonymize (full process out of scope – let's discuss, however)



Use software that performs & documents both data analysis process and data (such as R)

Anonymize*, record metadata, and select license

Publish in an open science repository in a FAIR way (such as Zenodo)

Use permanent identifiers and link with your research publications

* As open as possible, as closed as necessary (GDPR, commercial interests, vulnerable groups etc.)

FAIR revisited

FAIR principles applied https://about.zenodo.org/principles/

- Findable – metadata, discovery

Example: Zenodo metadata, searchable, unique identifier (DOI)

- Accessible – openly available

Example: Zenodo APIs, search exchanges

- Interoperable – data exchange, standards, vocabularies

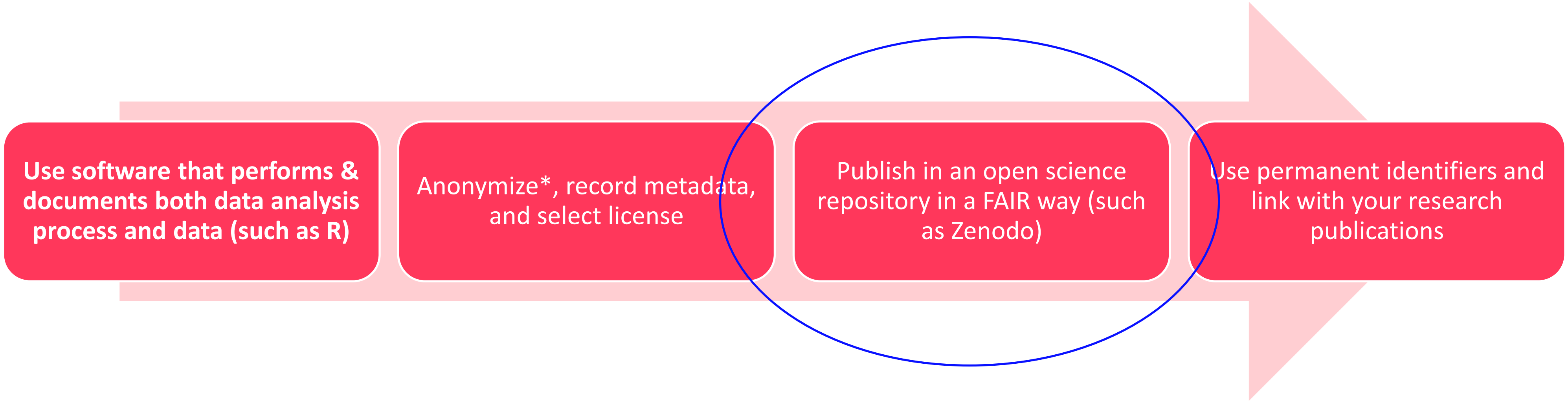
Example: csv format, metadata follows standard schema & exportable to Dublin Core, data processing software openly available

- Reusable – licensed to permit reuse and rich documentation

Example: RMarkdown documentation, creative commons license

Open data workflow: Step 3

Rstudio export and Zenodo demo



Use software that performs & documents both data analysis process and data (such as R)

Anonymize*, record metadata, and select license

Publish in an open science repository in a FAIR way (such as Zenodo)

Use permanent identifiers and link with your research publications

* As open as possible, as closed as necessary (GDPR, commercial interests, vulnerable groups etc.)

Rstudio export and Zenodo import demo

Combine data description, data analysis commands, and output

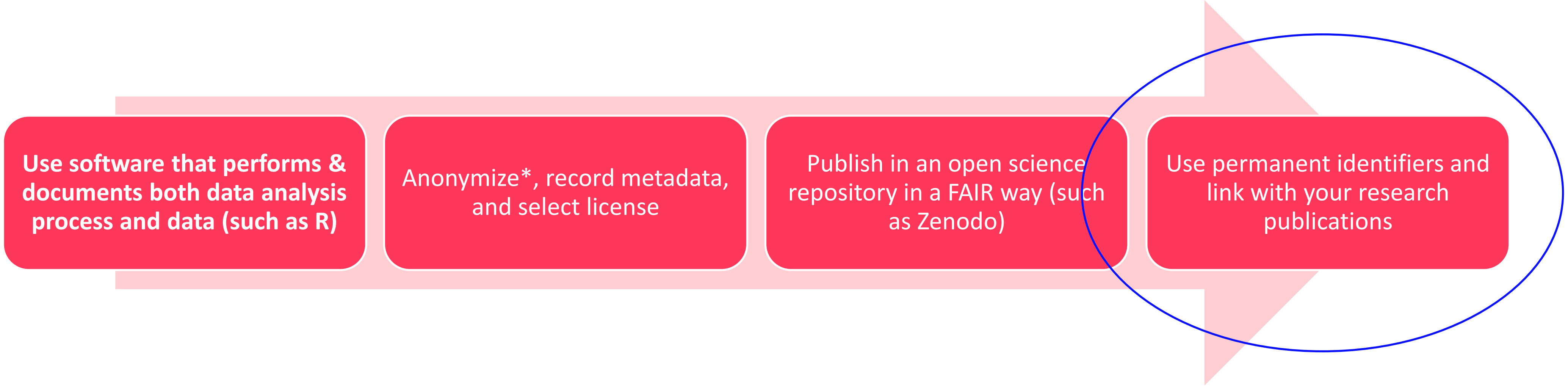
Two ways to do it:

1. Download as ZIP and upload to Zenodo (“bad” but efficient, good enough and workable for one-shot analysis)
2. Set up a Git repository (GitHub / GitLab), have versioning for your project, and use Zenodo connector to export versions (better, but… can be an overkill for solo or one-shot projects)

I will demo way 1 now and Natasha will demo way 2 later

Open data workflow: Step 4

Zenodo demo



Use software that performs & documents both data analysis process and data (such as R)

Anonymize*, record metadata, and select license

Publish in an open science repository in a FAIR way (such as Zenodo)

Use permanent identifiers and link with your research publications

* As open as possible, as closed as necessary (GDPR, commercial interests, vulnerable groups etc.)