D4.8 Report from thematic meeting I:
One Health EJP ASM Satellite Workshop on Data Management and Digital Innovation

Workpackage 4

Responsible Partner: SVA

Contributing partners: Sciensano, University of Surrey

GENERAL INFORMATION

European Joint Programme full title	**Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards**
European Joint Programme acronym	**One Health EJP**
Funding	**This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.**
Grant Agreement	**Grant agreement n° 773830**
Start Date	**01/01/2018**
Duration	**60 Months**

DOCUMENT MANAGEMENT

Deliverable	**D4.8 Open Data Access Point**
WP and Task	**WP4; Task 4.5.2**
Leader	**Sciensano (Valérie De Waele)**
Other contributors	**University of Surrey (Jade Passey, Piyali Basu, Roberto La Ragione) and SVA (Ann Lindberg)**
Due month of the deliverable	**M12**
Actual submission month	**M17 (workshop), M18 (report)**
Type ***R: Document, report*** ***DEC: Websites, patent filings, videos, etc.*** ***OTHER***	**OTHER and R**
Dissemination level ***PU: Public*** ***CO: confidential, only for members of the consortium (including the Commission Services)***	**PU**

ONE HEALTH EJP ASM SATELLITE WORKSHOP ON DATA MANAGEMENT AND DIGITAL INNOVATION

Table of Contents

A Programme

Annual Scientific Meetings (ASMs) are held within a One Health concept specifically showcasing the knowledge and scientific advances in Joint Research Projects and Joint Integrative activities. Alongside each ASM, a satellite workshop is run focusing on one of the priority areas for the OHEJP. The chosen theme for 2019 was '**Data Management and Digital Innovation'**.

The Satellite Workshop was held on Tuesday 21st of May at Teagasc institute in Dublin, Ireland. The workshop was announced by WP6 (communication) through various media, such as the website and the social media of One Health EJP. WP4 (integrative research) was in charge of the content of the workshop.

The detailed programme is presented below:

Satellite workshop on Data Management and Digital Innovation Programme

Time	Session	Speaker
12:00 – 13:00	*Lunch*	
13:00 – 13:10	Welcome, introduction	Ann Lindberg, WP4 and Roberto La Ragione, WP6
13:10 – 13:40	Introduction on OpenAIRE and European Open Science Cloud (EOSC) in health research	Niamh Brennan, Trinity College Library, Dublin
13:40 – 13:45	*Short break*	
13:45 – 15:00	Workshop: Planning your Data Management, part 1	Facilitators: Georgina Cherry, University of Surrey and Mickele Francisco, Sciensano
15:00 – 15:30	*Coffee break*	
15:30 – 16:45	Workshop: Planning your Data Management, part 2	Facilitators: Georgina Cherry, University of Surrey and Mickele Francisco, Sciensano
16:45 – 17:10	The value of Artificial Intelligence in tackling Foodborne Zoonoses (FBZ), Antimicrobial Resistance (AMR) and Emerging Threats (ET)	Nikos Papachristou, University of Surrey
17:10 – 17:25	Machine learning in digital pathology to improve cancer diagnosis in humans and animals	Ambra Morisi, University of Surrey
17:30	Close	

Conference Organisers:
Ann Lindberg- Work Package 4 Leader
Roberto La Ragione- Work Package 6 Leader
Valerie De Waele- Sciensano

B Audience

The satellite workshop was open to all audiences to attend who are working in the area of One Health. Priority was given to those who are registered as members of one of the OHEJP partner institutes. Thirty-five delegates attended the workshop (see Annex 1: Attendance list). This public was heterogene in regards to experience in data management, including early career researchers and senior researchers.

C Objective

Two interlinked themes were chosen for this workshop: 'Data Management and Digital Innovation'. Digital technologies are relatively new to the health innovation scene. Most scientific working in health domains want to participatein digital innovation, but they are inexperienced. Efforts to develop appropriate digital innovations for their institutions present critical organizational and multidisciplinary collaboration challenges for successful implementation.

The development of a data management plan is a key element for enabling institutional digital implementation and interinstitutional cooperation. Data management allows for new and more appropriate solutions in the life cycle of research data, which have an impact on the dissemination of the research data in a sustainable manner. Most of these solutions concern infrastructures and digital tools.

The workshop was designed to introduce researchers in 'Data Management and Digital Innovation' from both the European and institutional perspectives. It includes a practical and interactive exercise to get familiarized with the development of a data management. The satellite workshop provides also opportunities for early career researchers to present their research related to digital innovations.

D Content and speakers

To introduce the workshop, Ann Lindberg (SVA) and Roberto La Ragione (University of Surrey) welcomed the delegates and presented the programme of the workshop was divided in three sections:

- the first section introduced participants to 'Data Management and Digital Innovation' in both European and institutional perspectives;
- the second section was practical and, through in interactive exercise, the audience get familiarized with the different elements and steps required to develop a data management;
- the last section was focusing on digital innovations in research and provides the opportunities for early career researchers to present their research in this domain.

The first speaker of the workshop was Niamh Brennan, who is a Programme Manager for Research Informatics in Trinity College Library Dublin. In Trinity College, she works on research reporting, evaluation and impact; and she is responsible for the development of Trinity's Research Support System and its institutional repository, TARA (Trinity's Access to Research Archive). Niamh is also a member of several national and international groups working on open access to research outputs and enabling their improved reporting, retrieval and evaluation. These include Ireland's National Open Science Forum (which represents all Irish funding councils and research agencies and institutions), DART-Europe (Digital Access to Research Theses Europe), OpenAIRE2020 (Horizon2020) and OpenAIRE Advance (Horizon2020). Niamh Brennan presented the context of Research Data Management in terms of policies, rationale, requirements, infrastructure and supports. The latest developments in European, national and institutional policies were outlined along with the supports, tools and resources available to implement them e.g. the services provided by EOSC-hub, RDA and OpenAIRE Advance. She presented also a mapping of these services and supports to the research data lifecycle. In addition, her presentation provided information about the various requirements, wich fit together including GDPR, the Data Protection Impact Assessment (DPIA), Ethics Statements, Data Sharing Agreements, Data Management Plans (DMP), Dissemination/Impact Plans and reporting requirements. The slides of Niamh Brennan's presentation are available in the DMP group of the OHEJP website and also included in the present document.

To begin the practical section of the workshop, Georgina Cherry (University of Surrey) introduced delegates to the FAIR principles and discuss the benefits of creating a data management plan. The slides are in the annex of the present deliverable. Georgina has extensive experience in research data management. She has undertaken projects to improve search functionality by implementing natural language processing of search engine queries; building data structures using text mining and web analytics; importing digital content to a government information portal; and managing information systems, data privacy and user access. At the Veterinary Health Innovation Engine (vHive) research centre, Georgina is deriving actionable insights from data including the African Livestock Productivity and Health Advancement (ALPHA) Initiative.

The presentation of Georgina was followed by a description of the Data Management Plan (DMP) exercice. This description and the interactive animation of the practical section of the workshop was conducted by Mickele A.D. Francisco. Mickele is a computer scientist specialized in computer analysis and algorithmic. He currently leads innovation projects at the innovation lab using SAP new technologies & innovation services involving IoT, Big Data, Blockchain, Design Thinking, Machine Learning, Analytics and Low-code platform development such as Mendix. His role is to drive research, development and innovations using SAP new technologies, to increase internal and external client's productivity. He is specialized in Algorithm conception and process automation in a wide range of statistical languages and development tools. The interactive workshop was also facilitated by Georgina Cherry and Valérie De Waele (Sciensano). The delegates received either on paper or electronic format the material for the exercise, which is included in the annex of the deliverable and made available for other partners in the DMP group of the OHEJP website. The exercise objective was to develop a simple data management plan based on a practical scenario. This exercise includes a project scenario with research data and metadata, and the presentation of different tools useful to develop a data management plan, including checklists, templates and DMP examples. At the end of the practical exercise, Valérie De Waele presented DMPonline tool and the features of this tool, which facilitate finding information on the different elements of the DMP and facilitate collaboration between project partners in the writing process of the DMP.

These satellite workshops has provided opportunities for post-doctoral researchers (early career researchers) to present their research on digital topics. Both presentations were related to the use of Artificial Intelligence in health. Nikolaos Papachristou, University of Surrey, described the value and examples of "Artificial Intelligence in tracking Foodborne Zoonoses (FBZ), Antimicrobial Resistance (AMR) and Emerging Threats (ET)". Ambra Morisi, University of Surrey, presented his research in the development of a novel artificial intelligence method applied in veterinary histopathology to improve cancer diagnostic accuracy. The title of his presentation was "Deep Learning in Digital Pathology to improve cancer diagnosis in humans and animals".

E Annexes

1. Attendance List

Ambra Morisi
Andrew Byrne
Angela Lahuerta
Ann Lindberg
Claire Fitzgerald
Claudia Jaeckel
Dan Horton
Declan Murphy
Eithne Barron
Francesca Martelli
Georgina Cherry
Helen Brown
Irene Aldea
Jade Passey
Jens Andre Hammerl
Jose Gonzales
Karin Artursson
Maria Guelbenzu
Michaël Timmermans
Mickele Francisco
Niamh Brennan
Nikos Papachristou
Pikka Jokelainen
Pilar Pozo Pinol
Piyali Basu
Roberto La Ragione
Rosarie Lynch
Siobhán McCarthy
Sophie Granier
Taran Rai
Valerie De Waele
Vicente Lopez Chavarrias
Virginia Filipello
Tanel Tenson
Aidan Gonulez

2. Speakers' bio

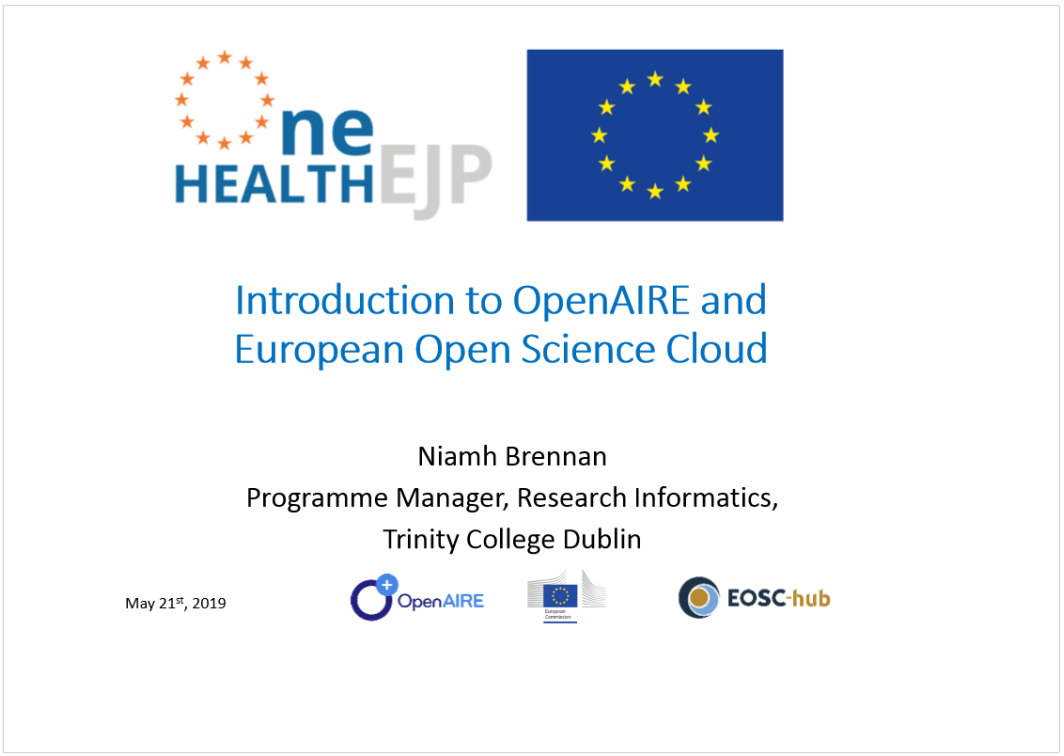
Niamh Brennan is Programme Manager for Research Informatics in Trinity College Library Dublin where she works on research reporting, evaluation and impact. She is responsible for the development of Trinity's Research Support System and its institutional repository, TARA (Trinity's Access to Research Archive). Niamh is a member of several national and international groups working on open access to research outputs and enabling their improved reporting, retrieval and evaluation. These include Ireland's National Open Science Forum (which represents all Irish funding councils and research agencies and institutions), DART-Europe (Digital Access to Research Theses Europe), OpenAIRE2020 (Horizon2020) and OpenAIRE Advance (Horizon2020). She is a member of the management councils of two key Irish journals in economics and social sciences and has partnered in a number of research projects in digital humanities, international development and social sciences. Niamh coordinated a national sectoral project in Ireland, funded by the HEA and managed by the IUA, which developed research reporting standards and research evaluation methodologies with a particular focus on developing new ways of demonstrating research impact. She has acted as a consultant on bibliometrics for HEA/Forfas (2011) and HEA (2106-2017). She is a member of the European Commission Expert Group on Skills for Open Science, reporting to the European Open Science Policy Platform (expert group report published: September 2017).

Georgina Cherry BSc MSc MCLIP is a chartered information professional who graduated from City University with a Master's degree in Information Science in 2004. Her information management career spans the higher education, financial services and technology sectors. She has undertaken projects to improve search functionality by implementing natural language processing of search engine queries; building data structures using text mining and web analytics; importing digital content to a government information portal; and managing information systems, data privacy and user access. The Veterinary Health Innovation Engine (vHive) is a research centre, startup and incubator supported by a co-investment of £8.5 million in resources dedicated to the development and adoption of new digital technologies in animal health. vHive utilises transformational digital and data analytics tools to advance the wellbeing of domestic animals through research, problem solving, education, training and knowledge brokering. At vHive, Georgina is deriving actionable insights from data including the African Livestock Productivity and Health Advancement (ALPHA) Initiative, a joint project funded by the Bill and Melinda Gates Foundation and Zoetis.

Mickele A.D. Francisco is a computer scientist specialized in computer analysis and algorithmic. He started his professional experience almost two decades ago working on SAP at IBM and since, worked on different technologies such in the field of nanotechnologies, where he was in charge of innovation, electron microscope automation, development and testing new approaches for automatic nanomaterial detection and classification. He currently leads innovation projects at the innovation lab using SAP new technologies & innovation services involving IoT, Big Data, Blockchain, Design Thinking, Machine Learning, Analytics and Low-code platform development such as Mendix. His role is to drive research, development and innovations using SAP new technologies, to increase internal and external client's productivity. He is specialized in Algorithm conception and process automation in a wide range of statistical languages and development tools such as Mendix for low-code, Spotfire for analytics, Python, Java, and other languages. He was also certified on the SAP Cloud Platform (SCP).

3. Presentations

Presentation 1




Carlos Moedas, Commissioner for Research, Science and Innovation
2017: '...opening up by default all scientific data produced by future projects under the €77 billion Horizon 2020 research and innovation programme, to ensure that the scientific community can re-use the enormous amount of data they generate.'
- http://europa.eu/rapid/press-release_IP-16-1408_en.htm


Grant Agreement Article 29.1– 6
"Each beneficiary must ensure open access to all peer-reviewed scientific publications"
"deposit research data … to make it possible for third parties to access, mine, exploit, reproduce and disseminate, free of charge"

All H2020 programmes & mandatory from January 2017

H2020 Data Management Requirements

Proposal stage: *These questions are asked on the proposal form*:

- What type of data will the project generate?
- What kinds of standards will be used?
- How will be data be shared & made available for verification & re-use? If the data cannot be made available, explain why.
- How will the data be curated and preserved?

Project stage: *The following will be required for all funded projects*:

- Initial Data Management Plan:6 months after project begins
- Mid-term Data Management Plan (DMP)
- Final DMP

Post project...

Funders Publication and Data Policies

Source:

● Full Coverage ◐ Partial Coverage ○ No Coverage

	Policy Coverage		Policy Stipulations					Support Provided			
Research Funders	Published outputs	Data	Time limits	Data plan	Sharing/ access	Long-term curation	Monitoring	Guidance	Repository	Data centre	Costs
AHRC	●	●	●	●	●	◐	◐	●	○	◐	●
BBSRC	●	●	●	●	●	●	●	●	●	◐	●
EPSRC	●	●	●	◐	●	●	●	◐	○	○	●
ESRC	●	●	●	●	●	●	●	●	●	●	◐
MRC	●	●	●	●	●	●	○	◐	●	○	◐
NERC	●	●	●	●	●	●	●	●	●	●	◐
STFC	●	●	●	●	●	●	●	◐	●	◐	◐
Cancer Research	●	●	●	●	●	●	●	◐	●	○	●
European Commission	●	●	◐	●	◐	◐	◐	●	●	◐	●
Wellcome Trust	●	●	●	●	●	●	●	●	●	●	●

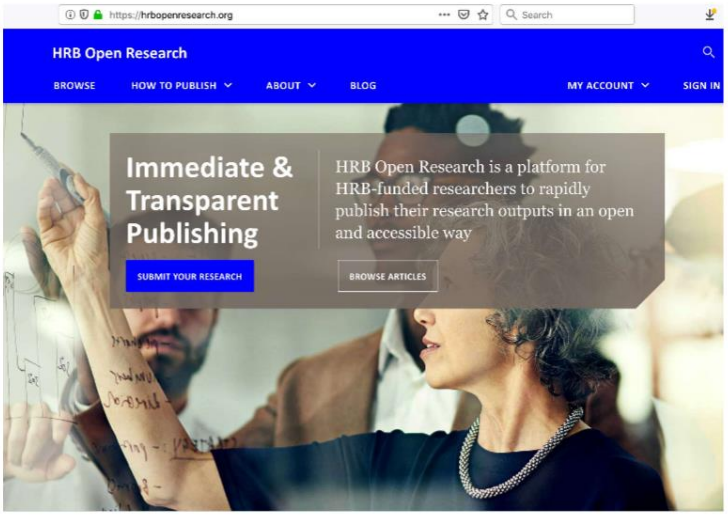
National Research Data Policies

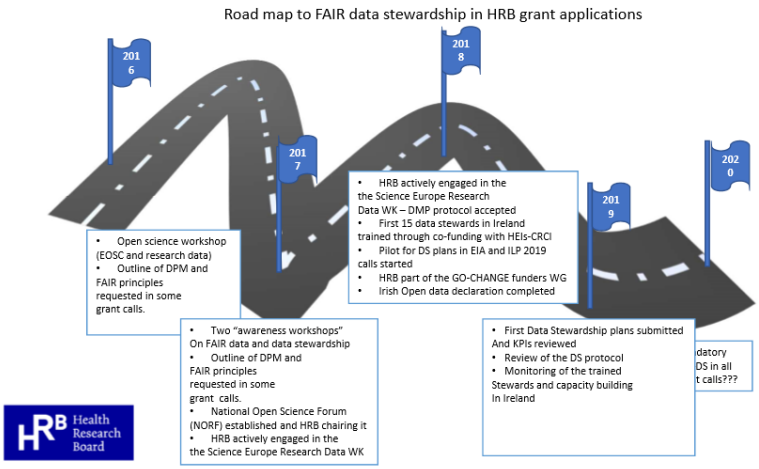
- 11 of 28 EU member states have national, research-data related policies in place
- The majority of these policies are owned by – or heavily involved – national research funders; and thus, generally focus on expectations for grant recipients.
- Policies generally took effect between 2009 and 2017 – with a clear upswing in activity during recent years

https://sparceurope.org/update-analysis-open-data-policies-finds-new-activity-around-oa-od-policies-multiple-countries/

SPARC Europe | DCC

An Analysis of Open Data and Open Science Policies in Europe, v2.1 (January 2018)





Science Europe Launches Framework for Discipline-specific Research Data Management

Research organisations and funders increasingly require researchers to create data management plans for their work and their proposals. Although researchers acknowledge the importance of good data management, creating such plans can be a time-consuming administrative affair. As there is no common standard format for these plans, they can be equally difficult to compare and evaluate for research organisations and funders. The Science Europe Working Group on Research Data has therefore developed a framework for the creation of domain-specific data management protocols.

- http://scieur.org/guidance-rdmps.

https://www.scienceeurope.org/wp-content/uploads/2018/12/SE_RDM_Practical_Guide_Final.pdf

June 2019

NORF
NATIONAL OPEN
RESEARCH FORUM

National Framework on the Transition to an Open Research Environment

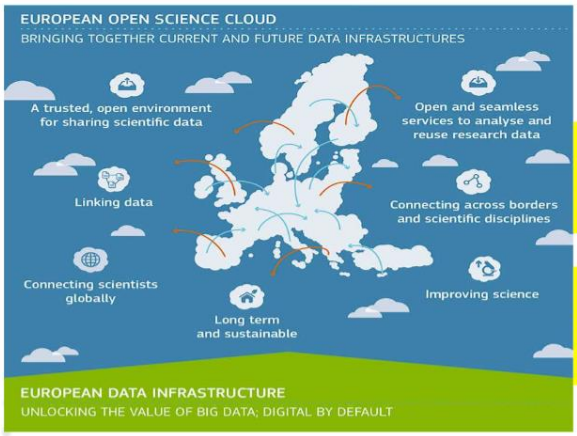
Preamble

The 'National Framework on the Transition to an Open Research Environment' has been developed as the first step in a process to create a National Action Plan for the transition to an Open Research environment in Ireland.

The 'Framework' is aligned with developing European Commission policy in this area and is structured accordingly. The European Commission Recommendation of 25 April 2018 on access to and preservation of scientific information asks Member States to 'set and implement clear policies (as detailed in national action plans)' covering: Open Access to Publications; Management of Research Data; Preservation and re-use of scientific information; Infrastructures for Open Research; Skills and Competencies; Incentives and Rewards.

The principles of this Framework support access to research funded by the Irish State. They support the free flow of information across national and international research communities, contributing to research-enabled teaching and learning, citizen science, open innovation, and greater transparency, accountability and public awareness of the results of publicly funded research. The transition to an Open Research environment has a key objective of enhancing and support of research excellence across all disciplines, research integrity, and public trust in research.

EOSC and EOSC-building projects



The EOSC is part of the overall European Cloud Initiative, which ultimately aims to connect business, industry and public facilities through the cloud.

EOSC-building projects are for instance
- OpenAIRE-Advance
- EOSC-hub
- EOSCpilot
- eInfraCentral
- FREYA

14

Project fact sheet EOSC-hub

- **Full title: Integrating and managing services for the European Open Science Cloud**
- **100 Partners, 76 beneficiaries (75 funded)**
- **3,874 PMs, 108 FTEs, more than 200 technical and scientific staff involved**
- **€33,331,180, funded by:**
  - European Commission: €30,000,000 (call H2020-EINFRA-2016-2017)
  - The participants of the EGI Foundation: €3,331,180
- **36 months: January 2018 – December 2020**

6/27/2019 15

EOSC-hub OpenAIRE

What is OpenAIRE?

1. Implement, monitor, align Open Science policies across Europe and the world
2. Harvesting of OA output, linking to contextual information
3. Deploy services to embed Open Science into researcher workflows
4. Develop global open standards for linking all research
5. Train for Open Science, for FAIR Science

16

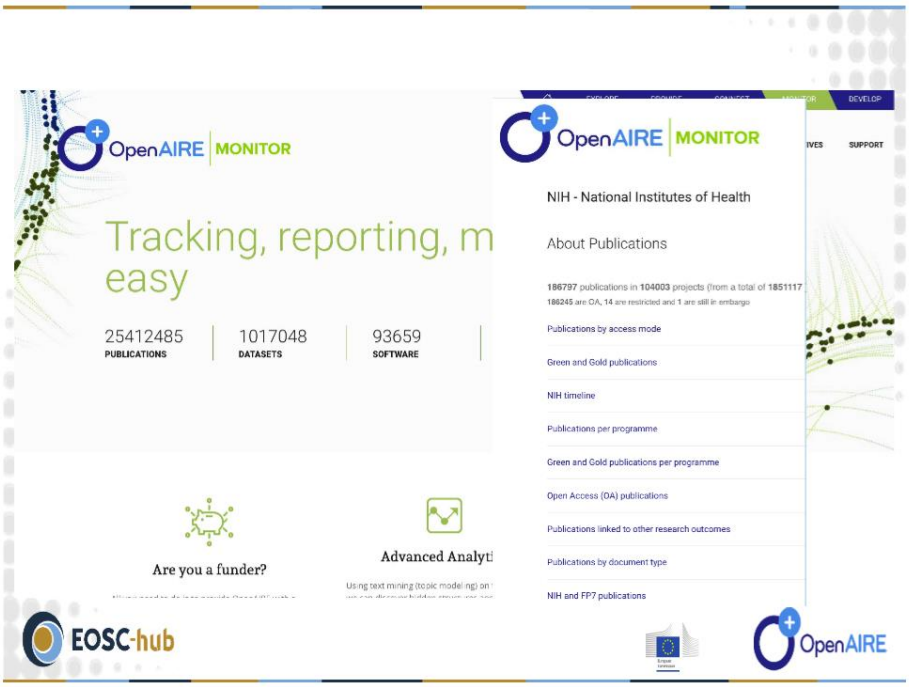
EOSC-hub OpenAIRE

OpenAIRE and its Services

https://www.openaire.eu/openaire-advance-project

- Integrated scientific information: links publication, project information, datasets... per funder/project/content provider...and presents them in an in one place
- Monitor and reporting on OS research outcomes for funders
- Training sessions and support on all subjects related to OS and OS policy
- Discovery of OS output per project, funder, data provider...
- Exchange of metadata and content amongst data providers
- A general purpose repository called Zenodo
- An Open Science helpdesk
- Monitoring and reporting mechanisms for research output per institution
- Analyses massive collections of documents, related meta-data and relational information

EOSC-hub OpenAIRE



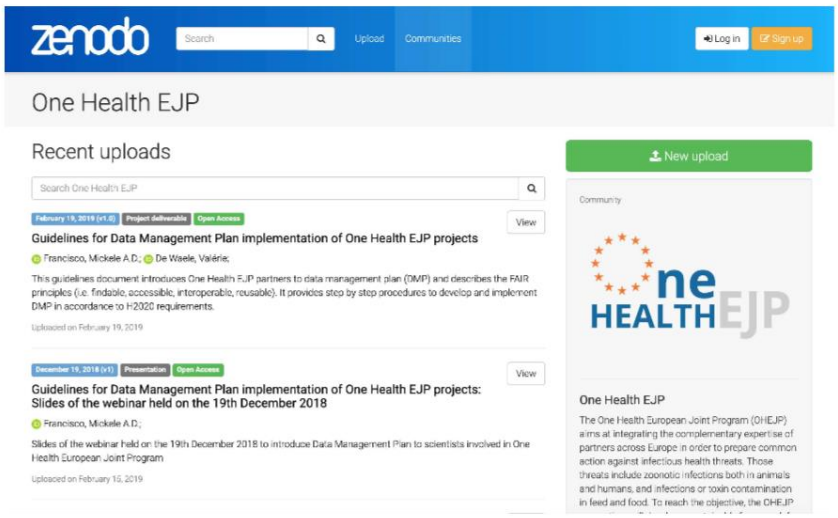
Short facts about Zenodo

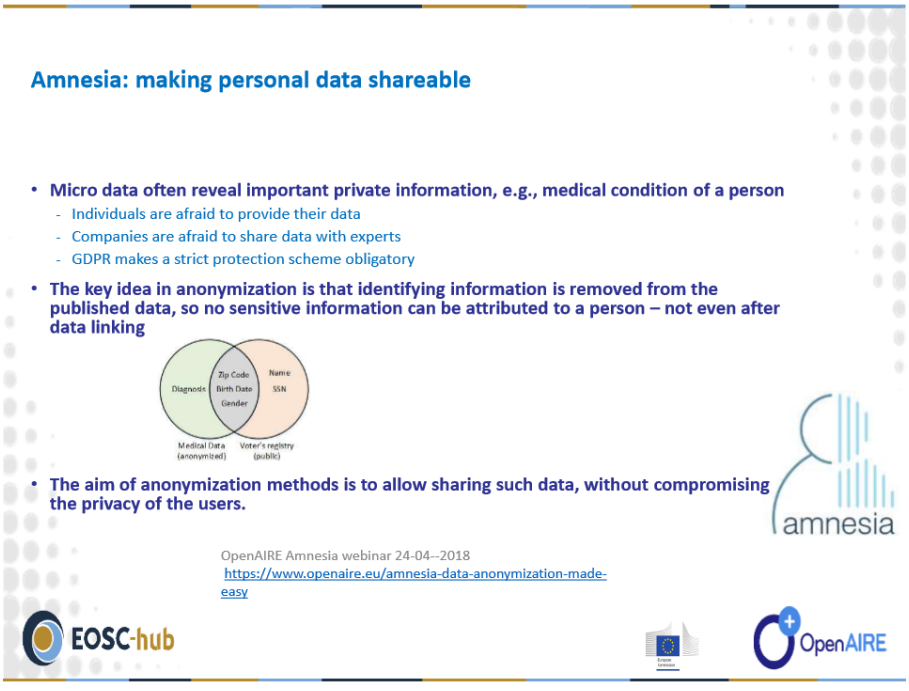
- Catch-all repository for EU-funded research
- Up to 50 GB per upload
- Data stored in the CERN Data Center
- Persistent identifiers (DOIs) for every upload, with DOI versioning
- Includes article-level metrics
- Free for the long tail of science
- Open to all research outputs from all disciplines
- GitHub integration
- Easily add EC funding information and report via OpenAIRE

Zenodo: https://zenodo.org/

DOI versioning in Zenodo

http://blog.zenodo.org/2017/05/30/doi-versioning-launched/


zenodo
Search
Upload
Communities
Log in
Sign up
One Health EJP
Recent uploads
Search One Health EJP
Guidelines for Data Management Plan implementation of One Health EJP projects
Francisco, Mickele A.D.; De Waele, Valérie;
This guidelines document introduces One Health EJP partners to data management plan (DMP) and describes the FAIR principles (i.e. findable, accessible, interoperable, reusable). It provides step by step procedures to develop and implement DMP in accordance to H2020 requirements.
Uploaded on February 19, 2019
View
Guidelines for Data Management Plan implementation of One Health EJP projects: Slides of the webinar held on the 19th December 2018
Francisco, Mickele A.D.;
Slides of the webinar held on the 19th December 2018 to introduce Data Management Plan to scientists involved in One Health European Joint Program
Uploaded on February 15, 2019
View
New upload
Community
One Health EJP
The One Health European Joint Program (OHEJP) aims at integrating the complementary expertise of partners across Europe in order to prepare common action against infectious health threats. Those threats include zoonotic infections both in animals and humans, and infections or toxin contamination in feed and food. To reach the objective, the OHEJP


Amnesia: making personal data shareable
Micro data often reveal important private information, e.g., medical condition of a person
Individuals are afraid to provide their data
Companies are afraid to share data with experts
GDPR makes a strict protection scheme obligatory
The key idea in anonymization is that identifying information is removed from the published data, so no sensitive information can be attributed to a person – not even after data linking
Diagnosis
Zip Code
Birth Date
Gender
Name
SSN
Medical Data (anonymized)
Voter's registry (public)
The aim of anonymization methods is to allow sharing such data, without compromising the privacy of the users.
amnesia
OpenAIRE Amnesia webinar 24-04--2018
https://www.openaire.eu/amnesia-data-anonymization-made-easy
EOSC-hub
OpenAIRE

Amnesia status

- Amnesia not only removes direct identifiers like names, social security numbers et cetera, but also transforms secondary identifiers like birth date and zip code so that individuals cannot be identified in the data.
- Amnesia is available as a public beta version at
  - https://amnesia.openaire.eu
- On-line version is for demonstration and testing purposes mostly (sample datasets included)
- Sensitive data can be anonymized locally by downloading the application
  - Security
  - Scalability
- OpenAIRE is in the process of adjusting it to health data, and looking for your feedback!
  - amnesia-helpdesk@imis.athena-innovation.gr

amnesia

OpenAIRE Amnesia webinar 24-04--2018
https://www.openaire.eu/amnesia-data-anonymization-made-easy

EOSC-hub OpenAIRE

EOSC-hub – OpenAIRE-Advance collaboration

- **Both in EINFRA-12 (topic A and B)**
  - EOSC-hub ~ storage, compute, application services
  - OpenAIRE ~ RDM; Publication services
- **Let's support Open Science together!**
  - Joint workplan plan
    - Technical integration of online services
    - Dissemination, community building, support, training
    - Governance

6/27/2019 24

EOSC-hub OpenAIRE

EOSC-Life

https://elixir-europe.org/news/eosc-life-start

https://www.eosc-portal.eu/eosc-life

SC-Life
oscLife

- 13 ESFRI research infrastructures in the Health and Food domain
- Create an open collaborative digital space for life science in the European Open Science Cloud (EOSC).
- Publish data from facilities and data resources in the EOSC & link these FAIR databases to open and reusable Tools and Workflows
- Accessible to users via Europe's national and international life-science clouds
- Connect users across Europe to a single login authentication and resource authorisation system
- develop data policies needed to preserve and deepen the trust given by research participants and patients volunteering their data and samples.

Establish EOSC-Life by publishing data and tools for cloud use

- Health and Food RI are **distributed with** many hundred partner facilities across Europe
- Health and Food RI have large community of tools developers across Europe

EOSC-Life

EOSC-Life


EOSC-Life
Project Duration: 1 March 2019 – 28 February 2023 (48 M)
Partners: 46 Partners, 17 Linked Third parties, 12 Third parties
Resources: 23.7 M€, 11 Work Packages, 34 Deliverables
CURATED DATABASES
SYSTEMS BIOLOGY
HIGHLY PATHOGENIC MICROORGANISMS
FUNCTIONAL GENOMICS
ADVANCED IMAGING TECHNOLOGIES
SCREENING AND MEDICINAL CHEMISTRY
STRUCTURAL BIOLOGY
MARINE MODEL ORGANISMS
TRANSLATIONAL RESEARCH
BIOBANKING
CLINICAL TRIALS
PLANT PHENOTYPING
MICROORGANISMS


Open and FAIR data

FAIR data principles

- **Findable**
  - Assign persistent IDs, provide rich metadata, register in a searchable resource, ...
- **Accessible**
  - Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...
- **Interoperable**
  - Use formal, broadly applicable languages, use standard vocabularies, qualified references...
- **Reusable**
  - Rich, accurate metadata, clear licences, provenance, use of community standards...

www.force11.org/group/fairgroup/fairprinciples
http://www.nature.com/articles/sdata201618

EOSC-hub OpenAIRE

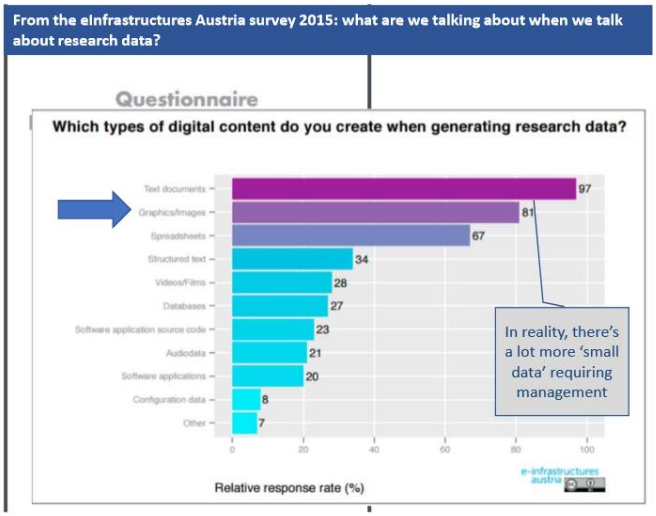
Principles =/= practice

H2020 DMP Guidelines: "This template is inspired by FAIR as a general concept."
Meaning: find your own (disciplinary) practice.

Guidelines on FAIR data management in Horizon 2020:
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

GO FAIR: initiative towards the internet of FAIR data and services. Started in Europe, but reaches out wide.
https://www.dtls.nl/fair-data/go-fair/
Infographic EC: http://ec.europa.eu/research/images/infographics/policy/open-data-2016-w920.png

EOSC-hub OpenAIRE



FAIR data ≠ open or accessible

Research data may be restricted permanently or for a specific for justifiable reasons, such as:

- Commercial exploitation
- Confidentiality
- Security
- Protection of personal data
- The achievement of the project's aim,
- Incompatibility with the further exploitation of the research results

– or other stated legitimate grounds. The important thing is to state whether some or all of it is restricted or not and give the reasons why. If any of it can be made openly accessible at any stage, similarly state this and how it will be effected.

https://5stardata.info/en/

Levels of Data Interoperability

5-star deployment scheme for open data

- Put your data on the web with an open licence. An open licence means that people can easily understand the terms under which your data is available for re-use.
- Make it available as structured data. For example, an Excel spreadsheet is more useable than a scan of a table in a PDF/DOC, and saves users from manually entering your data into their spreadsheet.
- Use open, standard formats. Non-proprietary formats can be accessed by any software - for example, you can save an Excel file as a CSV file, which is an open format.
- Use URIs to identify data. Uniform Resource identifiers (URIs) are a type of web link. They make it easier for your users to point at and draw upon your data.
- Link your data to other people's data. Linked data uses a common structure and format, so your data is standardised, and can more easily be joined with other datasets.

Excerpt from ict.gov.nz's Guide to the 5 Star Open Data Model licensed under CC BY 3.0 N

Services at the point of need

EOSC-hub and OpenAIRE services for a researcher?

Focus Clinic via Flickr cc

EOSC-hub

OpenAIRE

B2FIND

Making Open Science findable

http://b2find.eudat.eu/

Provided through EOSC-hub

- **Cross-disciplinary metadata and discovery service (B2FIND) allowing RI to make their data findable and discoverable in a central catalogue**
  - Metadata can be harvested via OAI-PMH. Possibility to use also APIs as JSON-API's and CSW2.0 to collect the metadata from the communities.
  - The project provides support to integrate community data catalogue

EUDAT

37

EOSC-hub OpenAIRE

B2DROP

Sync and share research data (https://www.eudat.eu/services/b2drop)

Provided through EOSC-hub:

- Store and share data with colleagues and team members, including research data not finalised for publishing
  - Cloud storage to share data with fine-grained access controls
  - Synchronise multiple versions of data across different devices, including workflow and computing environments
  - Publish data via B2SHARE

EUDAT

38

EOSC-hub OpenAIRE

B2SHARE

Store and publish data (https://b2share.eudat.eu/)

Provided through EOSC-hub:

- Data repository & publishing service (B2SHARE) allowing RIs to publish and manage data in a persistent way
  - Use of DataCite DOIs & EPIC PID
  - Domain specific metadata extensions
  - Manage the publish life cycle with version control
  - Community defined authorisation rules
  - Annotations via defined ontologies

EUDAT

39

EOSC-hub

OpenAIRE

B2SHARE - Public license selector

Choose a public license by answering some questions regarding access to your dataset.

Suggestions depend on several factors:

- Type of data
- Original licenses
- Data consumer access and distribution rights

Or use the search functionality.

40

EOSC-hub

OpenAIRE

B2NOTE

Use annotations to structure your data (https://b2note.eudat.eu/)

Provided through EOSC-hub:

- Manage and share annotations on data with colleagues and team members
  - Annotations are keywords or commentaries attached to a object, that explains or classifies it.
  - B2NOTE annotation service is integrated with the B2SHARE service and technology
  - B2NOTE can be easily integrated with other community data repository services
  - Provide training on semantic annotations

EUDAT

41

EOSC-hub OpenAIRE

Marketplace

Provided through EOSC-hub:

- **Marketplace: multi-tenant user-facing platform for service providers to publish their EOSC services and EOSC-compliant data repositories, and collect service orders**
  - Mature services and curated data
  - The RI retains control and accountability for the services and data published and participate in the management of the Hub service portfolio
  - Support to usage of common service templates

https://marketplace.egi.eu/

egi

42

EOSC-hub OpenAIRE

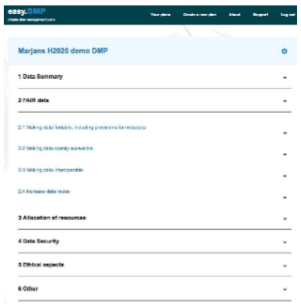
Complementary tools

EOSC-hub

OpenAIRE

Data management planning

- Recall that research funders like the EC and (academic) employers increasingly demand DMPs
- Tools available for writing your DMP



DMPOnline: https://dmponline.dcc.ac.uk/
EasyDMP: https://easydmp.sigma2.no/

EOSC-hub

OpenAIRE

DMP-writing tools

Both tools...

- ... contain the EC's Horizon2020 DMP template
- ... allow you to collaborate with others on your DMP (under construction)
- ... allow you to export your DMP
- ... plan to support "machine-actionable DMPs"

Guidance follows EC Guidance text more closely

Additional DCC guidance

Guidance is more interpretative

Pull-down menus to select e.g. metadata schema and file formats

Any feedback? support@easydmp.sigma2.no

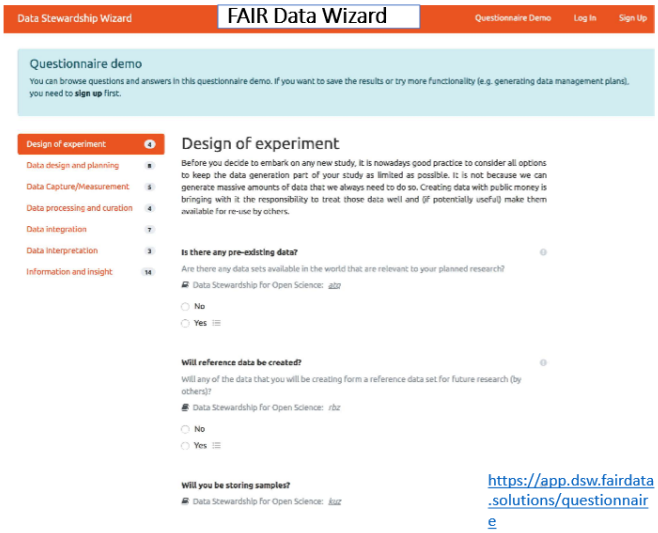
DMPOnline: https://dmponline.dcc.ac.uk/
EasyDMP: https://easydmp.sigma2.no/

EOSC-hub

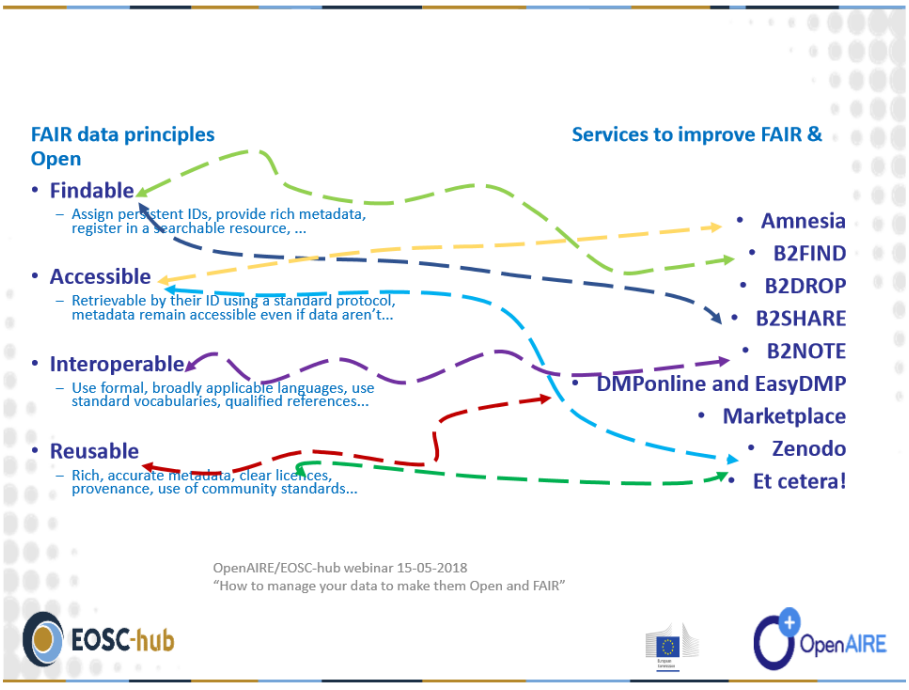
OpenAIRE

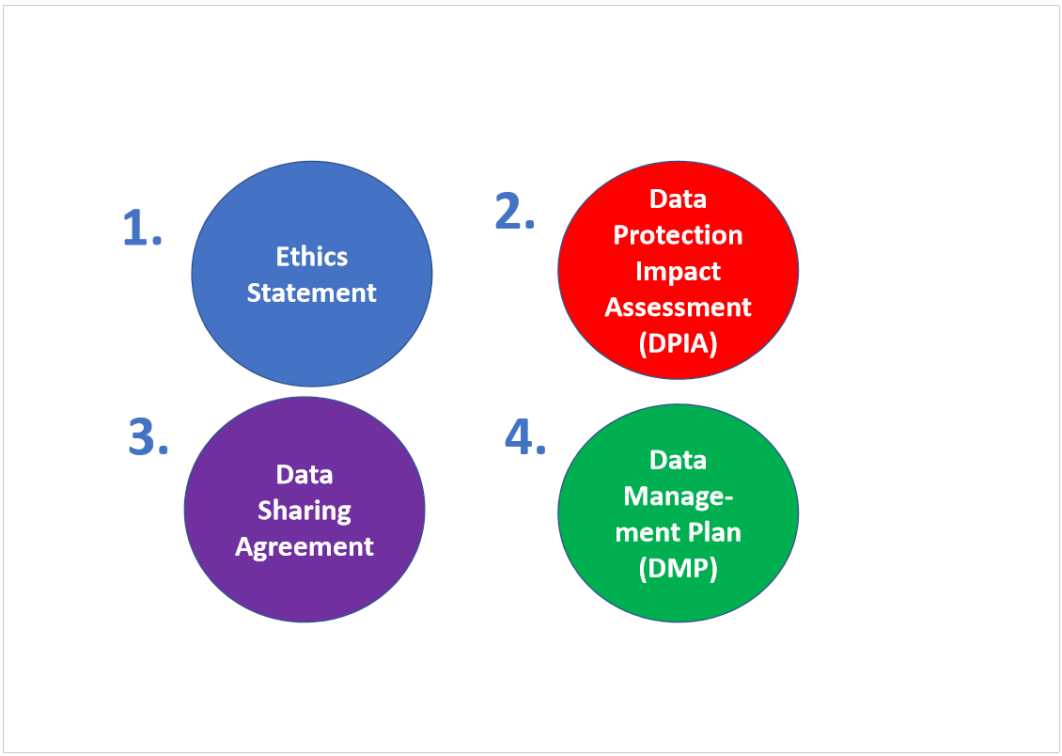
re3data.org

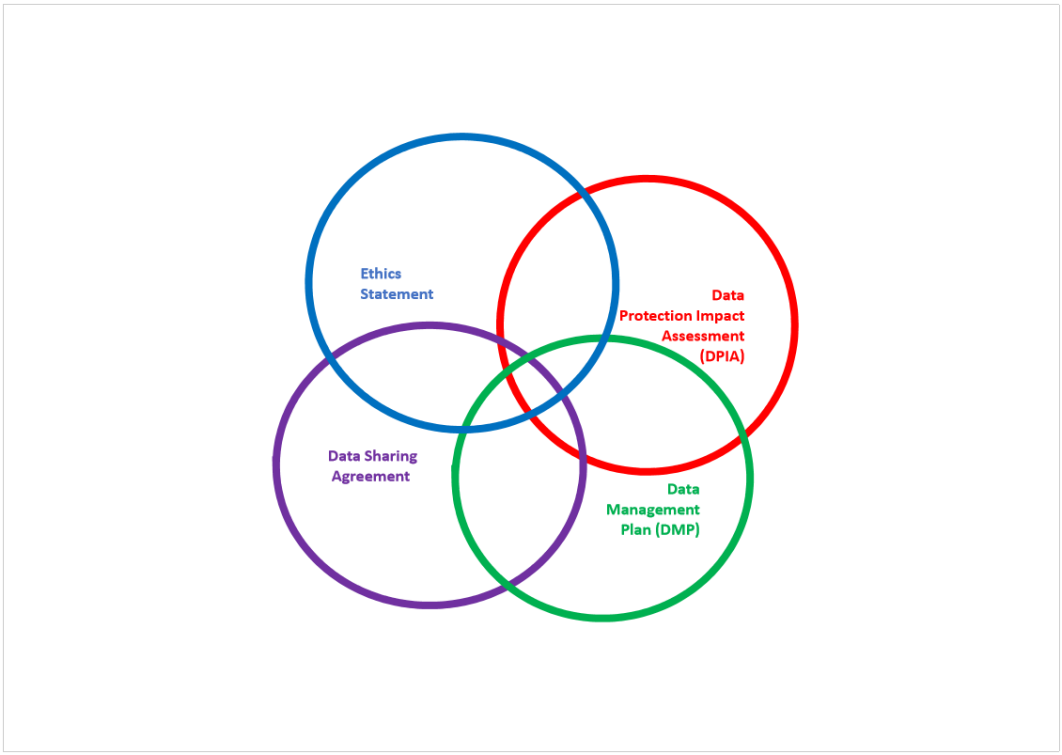
- Discovery tool for finding trusted repositories
- 1,254 in Life Sciences
- 568 in Medicine
- 177 in Agriculture



https://app.dsw.fairdata.solutions/questionnaire




1.
Ethics Statement
2.
Data Protection Impact Assessment (DPIA)
3.
Data Sharing Agreement
4.
Data Manage-ment Plan (DMP)


Ethics Statement
Data Protection Impact Assessment (DPIA)
Data Sharing Agreement
Data Management Plan (DMP)

Did you know?

- **When you integrate Open Science in your European research proposal, this makes your proposal more competitive.**
  - Grigorov, Ivo; Elbæk, Mikael; Rettberg, Najla; Davidson, Joy: "Winning Horizon 2020 with Open Science". https://doi.org/10.5281/zenodo.12247
- **There is evidence that grant proposals are receiving praise for including a DMP outline – even though in H2020 a DMP is not required at the proposal stage, and not a competitive point.**
- **Quotes from EC evaluation reviews of grant proposals:**
  - *"a clear description is provided of how core data sets and model development can be shared broadly within the scientific community"*
  - *"data storage and accessibility issues are not considered sufficiently"*
  - *"there is very good realization of the commercial potential of the project outcomes, which is reflected in the establishment of a data management plan, including IP related issues."*

Thanks to Ivo Grigorov (Technical University of Denmark, FOSTER project) for sharing these quotes.

Webinar May 14th 2018

EOSC-hub OpenAIRE

Impact

- Data citation
  - Data Citation Index (Clarivate)
  - Google Scholar
  - DataCite Statistics
- Link to Publications
- Highlighting via case studies
- Usage analytics
- Altmetrics
  - Impact Story
  - Figshare
  - Altmetric.com
  - PlumX

Image from ANDS: https://www.ands.org.au/__data/assets/pdf_file/0005/741740/data-impact-ebook.pdf

Acknowledgements

Slides re-used from the EOSC-hub and OpenAIRE-Advance projects, from Annalisa Montesanti (Health Research Board) and from 'How to manage your data to make them Open and FAIR', Ellen Leenarts (DANS) and Marjan Grootveld (DANS), May 15th 2018, https://www.slideshare.net/OpenAIRE_eu/20180515-how-to-manage-your-data-to-make-them-open-and-fair-public

Presentation 2

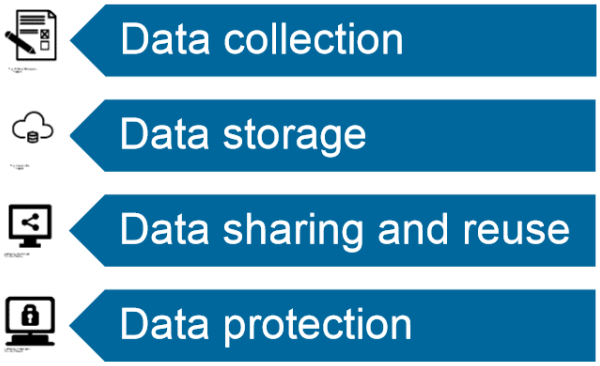
Key messages

Data management is **essential** to research

Creating a **data management plan** is a useful start

Effective data management aids **collaboration**

Data management is needed for…



What is a data management plan?

A data management plan is a formal document that describes what you will do with your data during the course of your research and after the project has been completed.

Benefits of good data management

- Planning upfront helps your project to succeed
- Encourages data reuse – ideas get bigger when they are shared
- Contingency plans for backing up and securely storing data
- Useful to include in your grant applications
- Increases the impact of your data
- Demonstrates integrity of researcher
- Tangible return on investment to share with your funders

FAIR principles

Findable
Accessible
Interoperable
Reusable

F - can others easily find and understand your data?

A - do people have to ask you for access to your data?

I - can others open your data files or do they need expensive software?

R - can others use your data to reproduce your results and verify your findings?

Why share your data?

Share with your future self – avoid repeating research!

- Promote your research – get cited!
- Enable new discoveries
- Replication
- Store your data in a reliable archive
- Comply with funding requirements

What do you need to share?

- Raw data
- Derived data
- Data underpinning publications
- Code
- Methods

What are research data in your context?

What would others need to understand your research?

Metadata

- Contextual information for data is called metadata — literally data about data
- Helpful for the reproduction of your work
- Should clearly and explicitly include the identifier of the data they describe
- Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process
- Since storing the metadata generally is much easier and cheaper than the data itself, finding the institution researcher or any needed information to track down the original information is still possible using only metadata
- So as a rule of thumb, you should never say ‘this metadata isn’t useful’; be generous and provide it anyway!

Policies for access, sharing and reuse

<< Data should be made available as soon as possible >>

- Obligations for sharing
  - Funding agency, institution
- Details of data sharing
  - How long?
  - When?
  - How to gain access to data?
- Ethical/privacy issues with data sharing
- Intended future uses/users
- Intellectual property and copyright issues
  - Institutional policies
  - Funding agency policies
  - Embargos for political or commercial reasons
- Citation
  - How should data be cited?
  - Persistent citation?

Restrictions on sharing your data

<< As open as possible, as closed as necessary >>

Are there privacy requirements from the funders or commercial partners?

- e.g. personal data, high security data

You might not have the right to share data collected from other sources

- Most databases are licensed and prohibit redistribution of data without permission

Discuss barriers to sharing your research data. These could be:

- Ethical, legal, professional
- Can these barriers be overcome? e.g. by anonymisation

Working with sensitive data

When working with research participants…

- Ensure you have obtained *valid consent*
- Pre-planning and agreeing with participants during the consent process, on what may and may not be recorded or transcribed, can be more effective than anonymisation
- Consider controlling access if anonymisation or consent for sharing are impossible

How to share your data

Deposit in a data repository eg. GenBank
- Funders' repository services

Online data sharing services
- Figshare, Zenodo, CKAN DataHub

Directories
- Re3data, DataBib

Data can be licensed under a CC-BY or CC0 declaration .

The **EUDAT B2SHARE** tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.

You may have rights to first use or to commercial exploit data

www.re3data.org EUDAT B2SHARE: https://b2share.eudat.eu/

Data retention and archiving

Where will it be preserved ?

How permanent are the data?

- Short term (e.g. 3-5 years)
- Long term (e.g. 10 years)
- Indefinite

What data will be preserved ?

Should discarded data be destroyed?

Who will be responsible ?

What are the re-processing costs?

Are there tools/software needed to create, process or visualise the data?

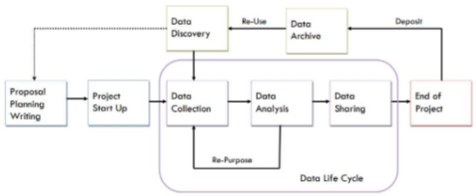
Research data lifecycle

Information has entropy and degrades over time due to:

- Researcher memory
- Storage misadventures
- Death of principal researcher etc.

Managing data in the research data lifecycle

- Choosing file formats
- File organisation and naming conventions
- Document all project/file details
- Access control and security
- Backup and storage
- Sharing and preservation



A good data management plan will cover

- What type of data or files will be produced or used?
- What standards will be used for documentation and metadata?
- What steps will be taken to protect privacy, security, confidentiality, intellectual property or other rights?
- If you allow others to reuse your data, how and when will the data be accessed and shared?
- How will the data be archived for preservation and long-term access?

Tools and resources to help create a plan

- DMP Guidance published by your institution
- DMPonline – tool providing step by step guidance to creating a data management plan (DMP)
- EasyDMP – online tool for creating a DMP
- Open Science Framework (OSF)

What is DMPonline?

- A web-based tool to help researchers write Data Management and Sharing Plans
- Includes requirements and guidance from funders, universities and other groups
- Developed by the Digital Curation Centre

Summary

Data management can
- Help you to be a more **efficient** researcher
- Increase the life-cycle of your data through **data sharing**

Create a **data management plan** before collecting any data
- There are online templates available at **DMPOnline**

Effective data management aids **collaboration**
- Partnerships create research opportunities

Presentation 3

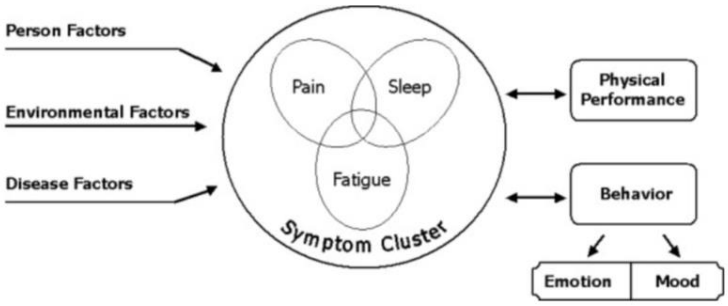
The value of Artificial Intelligence in tracking Foodborne Zoonoses (FBZ), Antimicrobial Resistance (AMR) and Emerging Threats (ET)

Nikolaos Papachristou
Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey

One Health EJP Annual Scientific Meeting Satellite Workshop 2019
21st May 2019, Dublin

Machine Learning & Symptoms

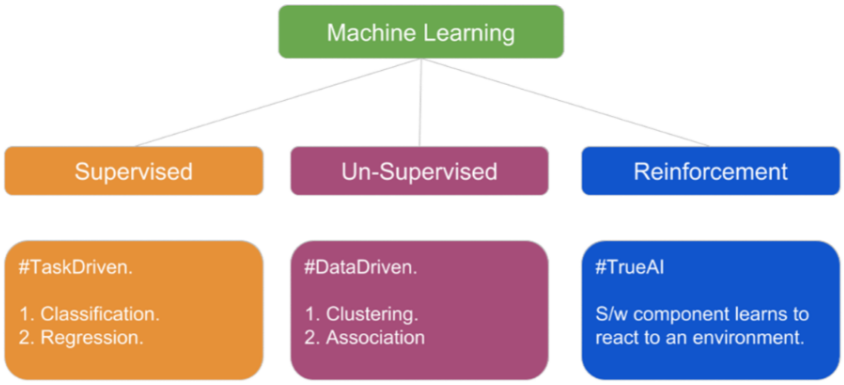
Symptom cluster in children and adolescents with cancer



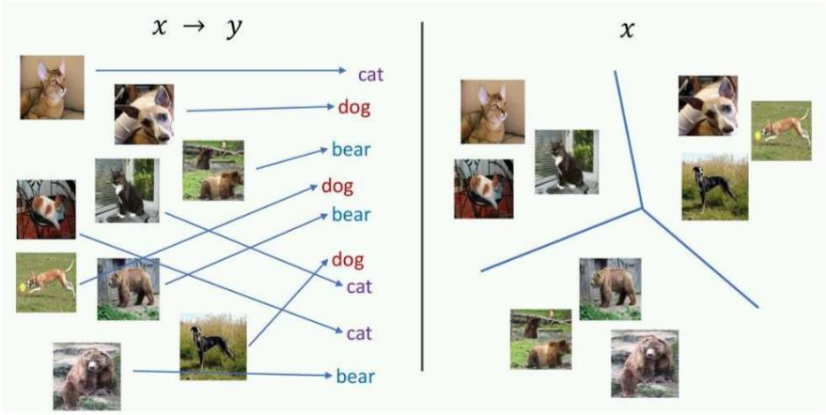
source: Hockenberry et al. "Symptom Clusters in Children With Cancer", 2007

Artificial Intelligence (AI)

Types of Machine Learning



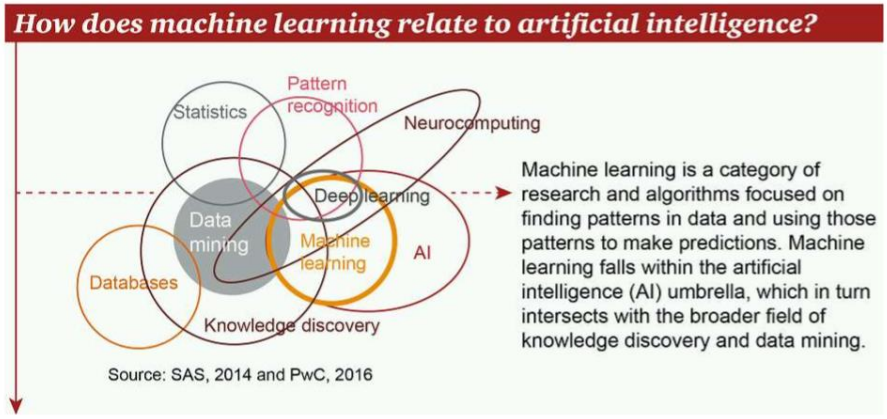
Supervised vs Unsupervised Machine Learning



Reinforcement Machine Learning

AI, Data Science, Machine Learning, Data Mining, Statistics, etc?

UNIVERSITY OF SURREY



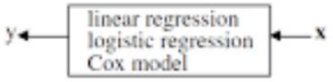
Machine learning overview (infographic), PWC, 2016

Statistical Science
2001, Vol. 16, No. 3, 199–231

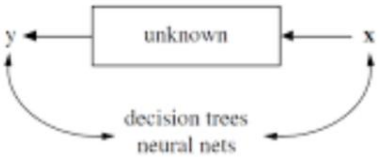
UNIVERSITY OF SURREY

Statistical Modeling: The Two Cultures

The Data Modeling Culture

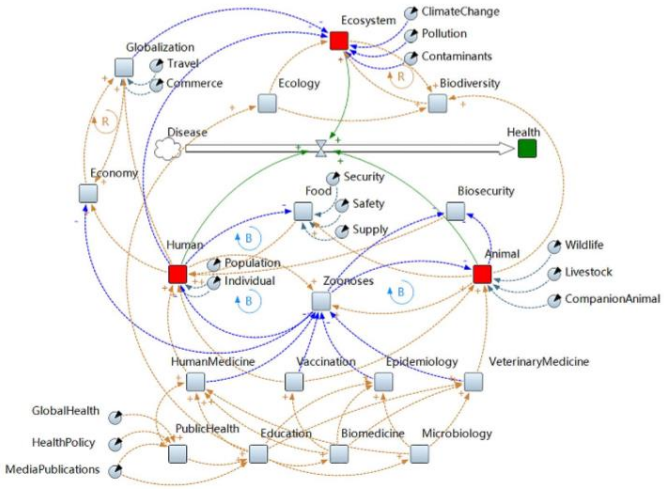


The Algorithmic Modeling Culture



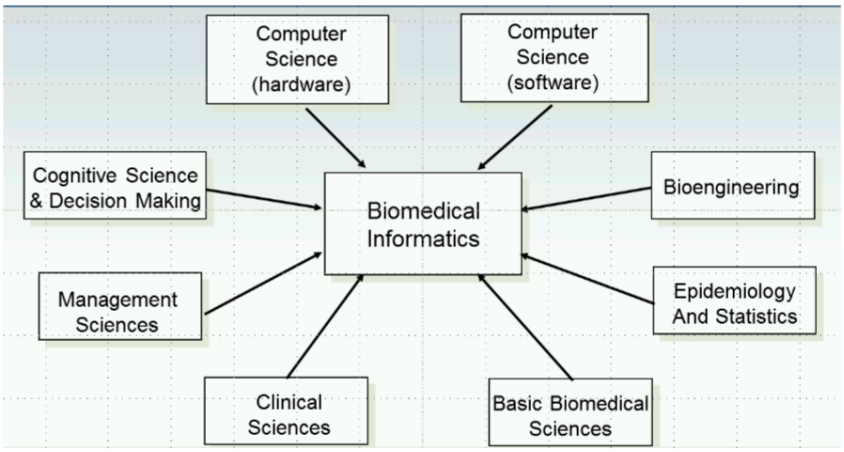
One Health & Informatics

The One Health concept



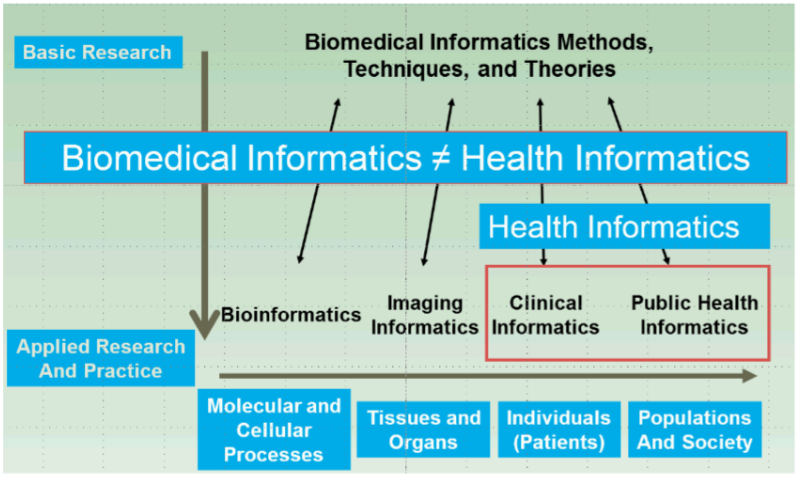
Xie, T et al. 2017

Veterinary Informatics?



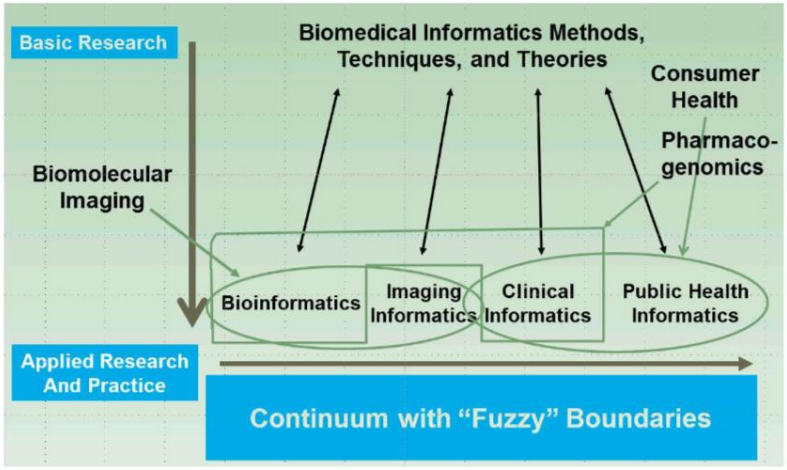
American Medical Informatics Association (AMIA)

AI & One Health methods?



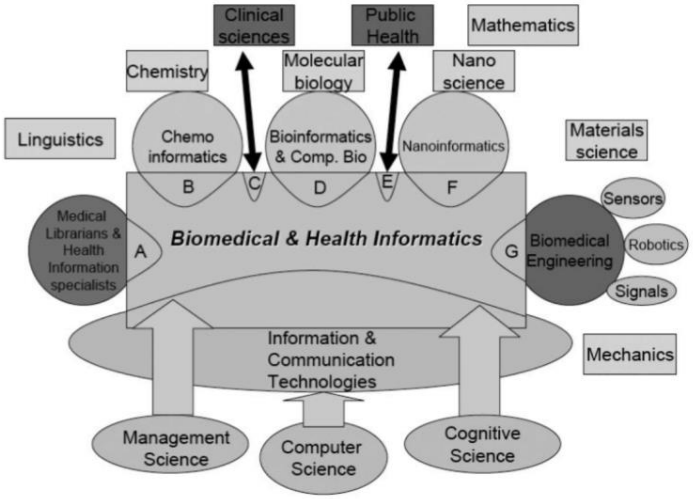
American Medical Informatics Association (AMIA)

AI & One Health Perspective?



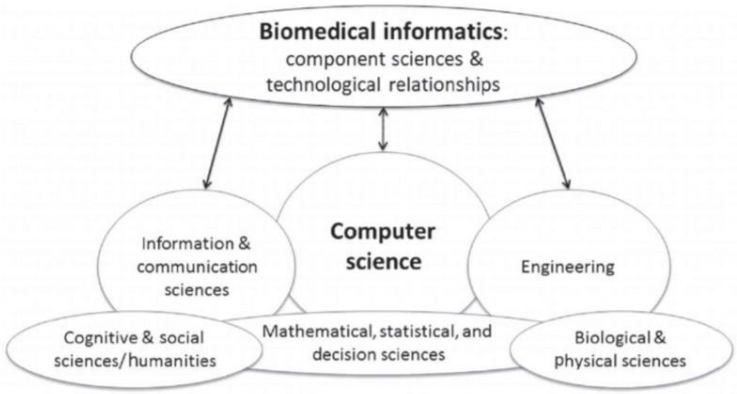
American Medical Informatics Association (AMIA)

AI & One Health related fields?



American Medical Informatics Association (AMIA)

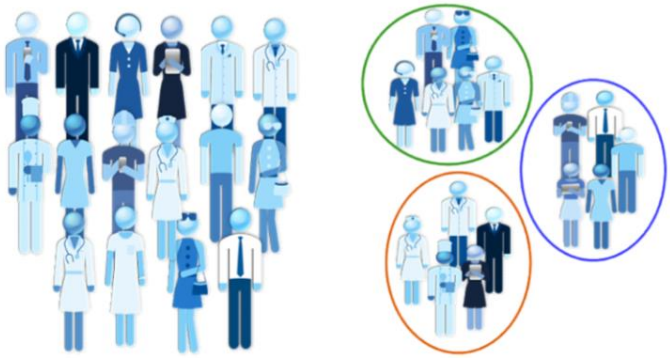
AI & One Health disciplines?



American Medical Informatics Association (AMIA)

Examples

1st: Clustering Cancer Patients

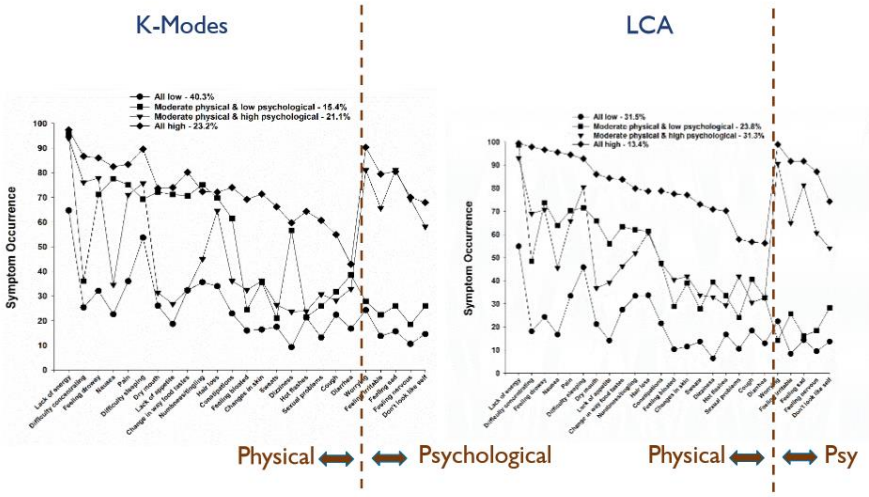


1. Nikolaos Papachristou et al. "***Comparing Machine Learning Clustering with Latent Class Analysis on Cancer Symptoms' Data***", in Proc. of the IEEE-NIH 2016 Special Topics Conference on Healthcare Innovations and Point-of-Care Technologies, November 2016.

2. Nikolaos Papachristou et al. "***Congruence Between Latent Class and K-modes Analyses in the Identification of Oncology Patients with Distinct Symptom Experiences***", Journal of Pain and Symptom Management, 2017

image: https://towardsdatascience.com/k-means-clustering-identifying-f-r-i-e-n-d-s-in-the-world-of-strangers-695537505d

k-Modes vs LCA



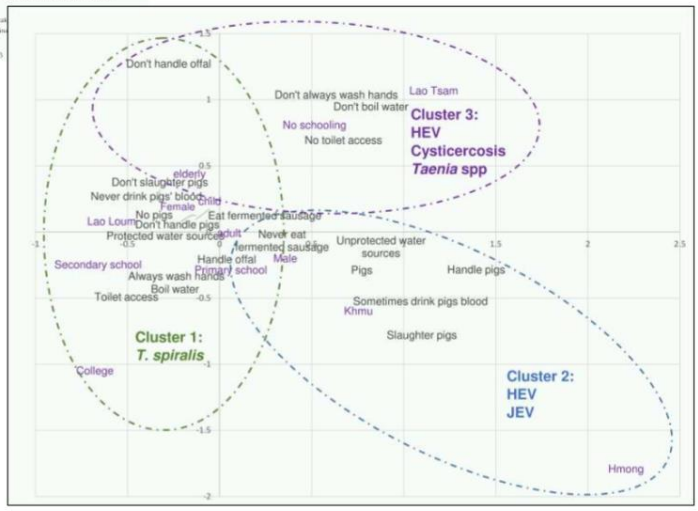
Pairwise comparisons for MSAS and QOL scales

Table 5: Differences in Memorial Symptom Assessment Scale (MSAS Summary Scores) inside each identified group of patients with k-Modes

MSAS Scores	Total n=1329 Mean (SD)	Cluster A n = 261 (45%) Mean (SD)	Cluster B n = 126 (21.7%) Mean (SD)	Cluster C n = 193 (33.3%) Mean (SD)	Cluster C n = 193 (33.3%) Mean (SD)	Statistics
Total number of symptoms	12.51 (6.31)	6.72 (3.15)	13.93 (2.84)	13.66 (2.84)	20.54 (4.05)	F(3,1325) = 1187.4 P < 0.001
MSAS PSYCH subscale score	0.89 (0.72)	0.37 (0.33)	0.58 (0.41)	1,27 (0.52)	1.64 (0.65)	F(3,1325) = 553.729 P < 0.001
MSAS PHYSICAL subscale score	0.8 (0.56)	0.38 (0.28)	1.08 (0.37)	0.74 (0.36)	1.39 (0.55)	F(3,1325) = 478.279 P < 0.001
MSAS Global Distress Index	1.01 (0.7)	0.44 (0.32)	1.01 (0.42)	1.25 (0.46)	1.78 (0.69)	F(3,1305) = 588.214 P < 0.001
MSAS Total Score	0.72 (0.48)	0.33 (0.2)	0.78 (0.26)	0.78 (0.26)	1.33 (0.44)	F(3,1325) = 765.761 P < 0.001

PLOS | NEGLECTED TROPICAL DISEASES

Endemicity of Zoonotic Diseases in Pigs and Humans in Lowland and Upland Lao PDR: Identification of Socio-cultural Risk Factors

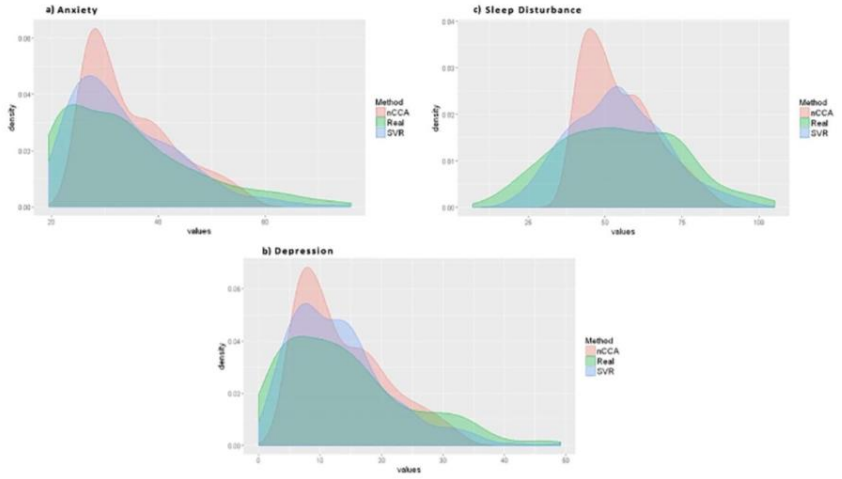


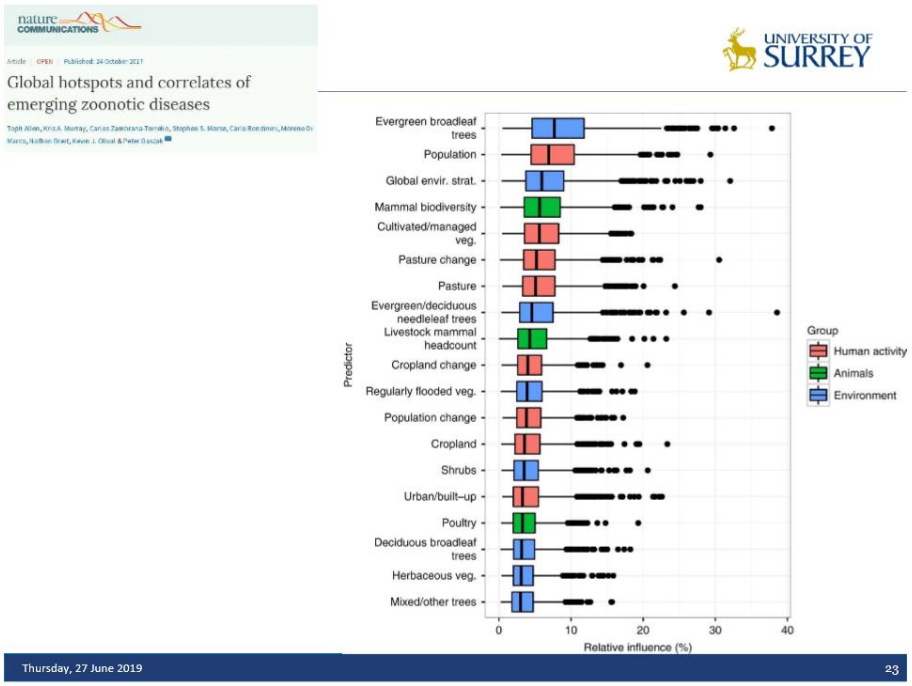
2nd: Predicting cancer symptoms (depression, anxiety, sleep disturbance)

Variables	Type of variable	Range
Age	Continuous	19.9 - 90.72
Gender	Nominal	1 - 3
Education	Continuous	4 - 23
BMI	Continuous	15.21 - 54.58
Ethnicity	Nominal	0-3
Married	Binary	0 - 1
Do you live alone	Binary	0 - 1
Working	Binary	0 - 1
Income	Nominal	1 - 4
Caregiver to children	Binary	0 - 1
Caregiver to adult	Binary	0 - 1
Karnofsky Performance status	Continuous	30 - 100
Self-administered comorbidity questionnaire (SCQ) score	Nominal	0 - 21
Number of metastatic sites out of 9	Nominal	0 - 6
Type of cancer	Nominal	1 - 4
Time of diagnosis to start of study in years	Continuous	0.041 - 38.32
Number of prior treatments out of 9	Nominal	0 - 7
Hemoglobin (HgB1)	Continuous	6.7 - 16.1
Exercise on a regular basis	Binary	0 - 1
Cycle length	Nominal	1 - 3
General Sleep Disturbance Scale	Continuous	7-119
Morning fatigue measured as part of the Lee Fatigue Scale	Continuous	0 - 9.84
Evening fatigue measured as part of the Lee Fatigue Scale	Continuous	0 - 10
Morning energy measured as part of the Lee Fatigue Scale	Continuous	0 - 10
Evening energy measured as part of the Lee Fatigue Scale	Continuous	0 - 10
Attentional function index	Continuous	0.54 - 10
Center for Epidemiological Studies-Depression Scale	Continuous	0 - 56
State Anxiety Scale	Continuous	20 - 80
Occurrence of pain	Nominal	0 - 2

1. Nikolaos Papachristou et al. "***Learning from Data to Predict Future Symptoms of Oncology Patients***", PLoS ONE 13(12), 2018.

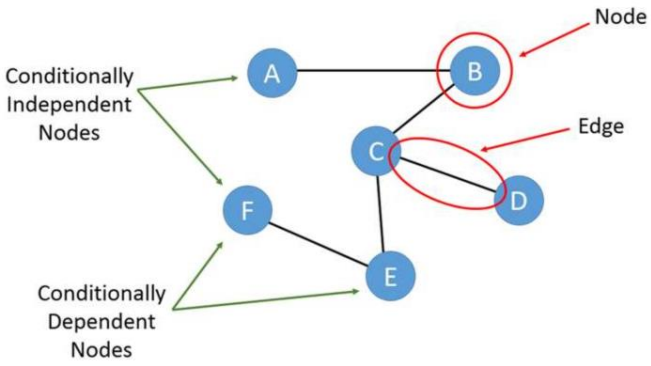
Predicting cancer symptoms (depression, anxiety, sleep disturbance)





3rd: Network Analysis of the Multidimensional Symptom Experience of Oncology

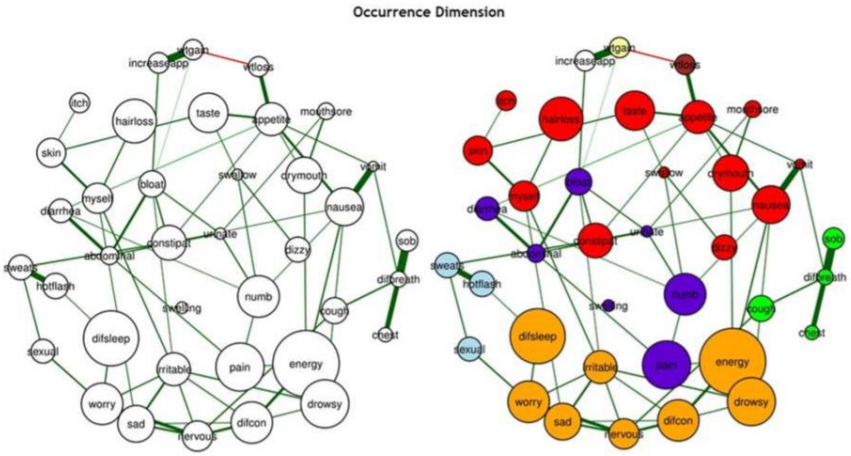
UNIVERSITY OF SURREY



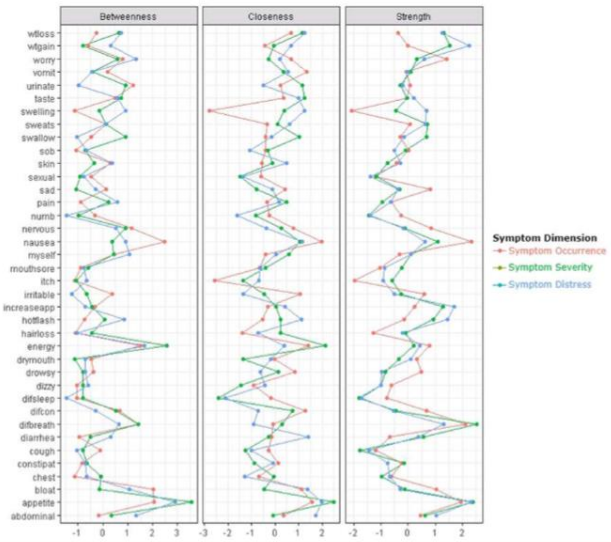
1. Nikolaos Papachristou et al. "***Network Analysis of the Multidimensional Symptom Experience of Oncology***", Scientific Reports, Nature, volume 9, Article number: 2258, 2019.

Thursday, 27 June 2019 24

Network Analysis of the Multidimensional Symptom Experience of Oncology



Network Analysis of the Multidimensional Symptom Experience of Oncology



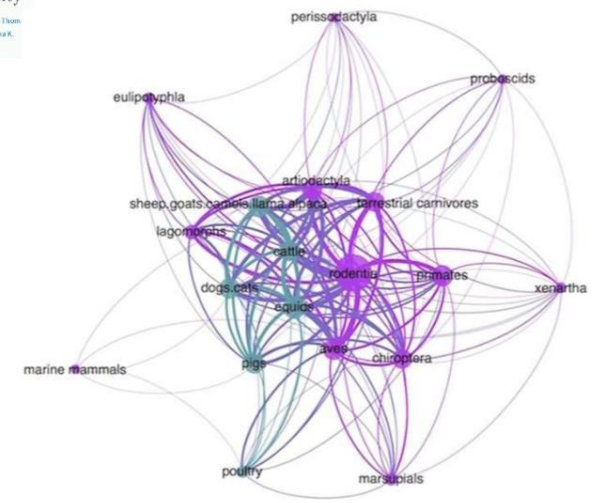
SCIENTIFIC REPORTS

Article | OPEN | Published: 07 October 2015

Spillover and pandemic properties of zoonotic viruses with high host plasticity

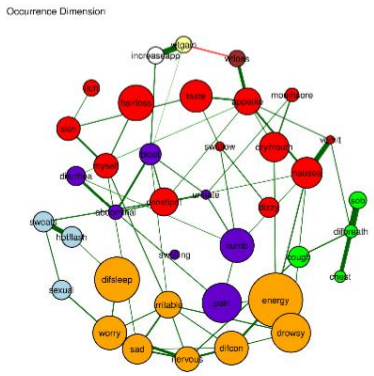
Christine Kreuder Johnson, Peta L. Hitchens, Tierra Smiley Evans, Tracey Goldstein, Kate Thorn, Andrew Clements, Damien O. Joly, Nathan D. Wolfe, Peter Daszak, William D. Karesh & Jonna K. Mazet

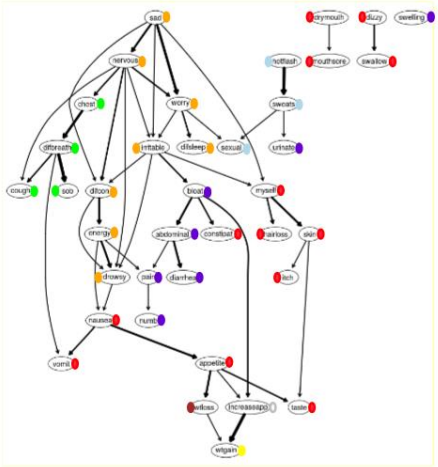
UNIVERSITY OF SURREY

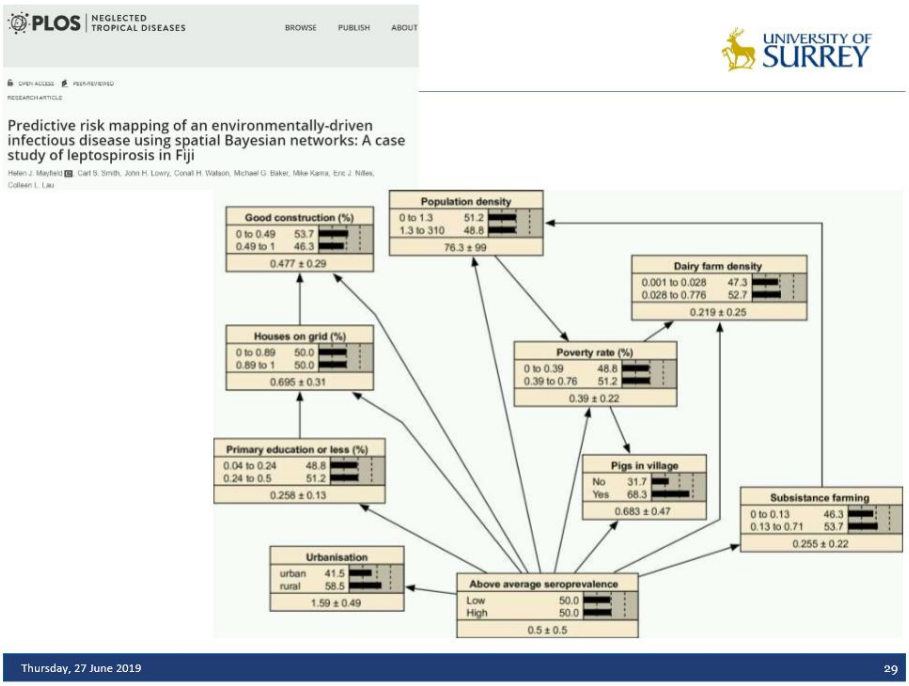


4th : Symptom sequential causality

UNIVERSITY OF SURREY






PLOS | NEGLECTED TROPICAL DISEASES
BROWSE PUBLISH ABOUT
RESEARCH ARTICLE
Predictive risk mapping of an environmentally-driven infectious disease using spatial Bayesian networks: A case study of leptospirosis in Fiji
Helen J. Mayfield, Carl S. Smith, John H. Lowry, Conall H. Watson, Michael G. Baker, Mike Kama, Eric J. Nilles, Colleen L. Lau
UNIVERSITY OF SURREY
Good construction (%)
0 to 0.49 53.7
0.49 to 1 46.3
0.477 ± 0.29
Population density
0 to 1.3 51.2
1.3 to 310 48.8
76.3 ± 99
Dairy farm density
0.001 to 0.028 47.3
0.028 to 0.776 52.7
0.219 ± 0.25
Houses on grid (%)
0 to 0.89 50.0
0.89 to 1 50.0
0.695 ± 0.31
Poverty rate (%)
0 to 0.39 48.8
0.39 to 0.76 51.2
0.39 ± 0.22
Primary education or less (%)
0.04 to 0.24 48.8
0.24 to 0.5 51.2
0.258 ± 0.13
Pigs in village
No 31.7
Yes 68.3
0.683 ± 0.47
Subsistance farming
0 to 0.13 46.3
0.13 to 0.71 53.7
0.255 ± 0.22
Urbanisation
urban 41.5
rural 58.5
1.59 ± 0.49
Above average seroprevalence
Low 50.0
High 50.0
0.5 ± 0.5
Thursday, 27 June 2019 29


CVSSP
Payam Barnaghi
Professor of Machine Intelligence
CVSSP, University of Surrey
Dr Shirin Enshaeifar
Andreas Markides
Tarek Elsaleh
Severin Skillman
Nikolaos Papachristou
Narges Pourshahrokhi
Roonak Rezvani
Honglin Li
Prof Christine Miaskowski, UCSF
Dr Maria Bermudez-Edo, Uni Granada
Dr Frieder Ganz, Adobe
Tuesday, 21 May 2019 30

Thank you very much

University of Surrey: https://www.surrey.ac.uk/people/nikolaos-papachristou

email: n.papachristou@surrey.ac.uk

Deep Learning in Digital Pathology to improve cancer diagnosis in humans and animals

Ambra Morisi[1] & Taran Rai[2]

N. J. Bacon [4], S. A. Thomas[3], M. Bober[2], K. Wells[2], B. Bacci[5], R. La Ragione[1]

[1]School of Veterinary Medicine, University of Surrey, Guildford, GU2 7AL, United Kingdom

[2]Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, United Kingdom

[3]National Physical Laboratory, Teddington TW11 0LW, United Kingdom

[4]Fitzpatrick Referrals Oncology and Soft Tissue, Guildford, United Kingdom

[5]Department of Veterinary Medical Sciences, University of Bologna, Bologna, Italy

Cancer

- High incidence and mortality rate
- Cancer diagnosis through histopathology

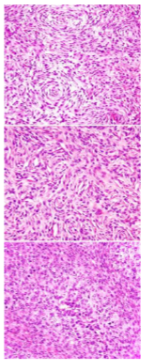
Tumour selection for the project

Tumour selected for the project:
- **Canine Soft Tissue Sarcomas (STSs)**

Other samples:
- Canine Mast Cell Tumours (MCTs)
- Canine Anal Sac Adenocarcinomas (AGACAs)

Why?
- Many samples available
- Most common tumours
- Histologic grade

Canine Soft Tissue Sarcoma



Grade I

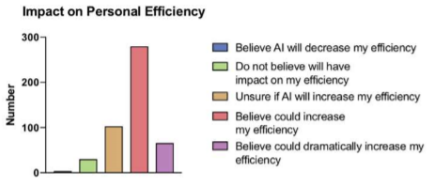
Grade II

Grade III

In human STS, there was a 75% agreement between 15 pathologists when 25 cases were evaluated with the STS grading scheme (*Coindre et al.*, 1986)

Advantages of Artificial Intelligence (AI) methods and Digital Pathology

- Increase the quality and accuracy of the diagnosis
- Increase the speed of execution
- Less Biased
- Operational Ability
- Improve work flow



Sarwar, Shihab, et al. "Physician perspectives on integration of artificial intelligence into diagnostic pathology." npj Digital Medicine 2.1 (2019): 28.

Overall aim of PhD project

Development of a novel artificial intelligence method applied for the first time in veterinary histopathology to improve cancer diagnostic accuracy

Challenges using AI in Digital Pathology

- **Low volumes of available data; less than 10000 samples**

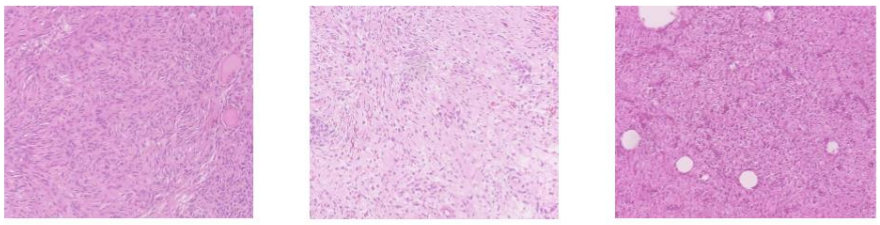
Challenges using AI in Digital Pathology

- Low volumes of available data; less than 10000 samples
- **Ground truth generation often expensive and unavailable**

Challenges using AI in Digital Pathology

- Low volumes of available data; less than 10000 samples
- Ground truth generation often expensive and unavailable
- **Stain variation and artefacts**

Stain variation and Artefacts (10x)



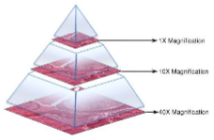
Grade 1 Grade 2 Grade 3

Challenges using AI in Digital Pathology

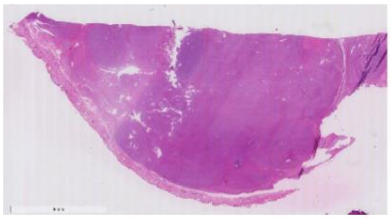
- Low volumes of available data; less than 10000 samples
- Ground truth generation often expensive and unavailable
- Stain variation and artefacts
- **Deep Learning models are considered “black boxes”**

Challenges using AI in Digital Pathology

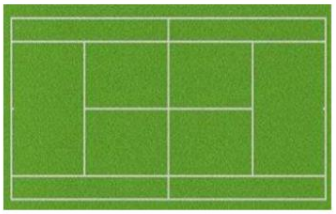
- Low volumes of available data; less than 10000 samples
- Ground truth generation often expensive and unavailable
- Stain variation and artefacts
- Deep Learning models are considered “black boxes”
- **WSIs are gigapixel images**



Gigapixel Whole Slide Image

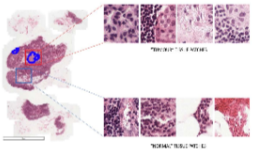


=

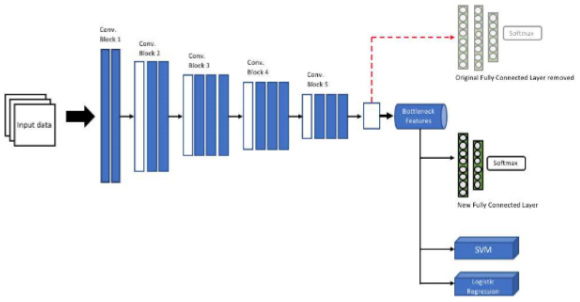


(Printed at 300 Dots Per Inch)

A Typical Deep Learning Architecture



Patches extracted from 60 WSIs (CAMLEYON)



Rai, T., et al. "Can ImageNet feature maps be applied to small histopathological datasets for the classification of breast cancer metastatic tissue in whole slide images?." *Medical Imaging 2019: Digital Pathology*. Vol. 10956. International Society for Optics and Photonics, 2019.

Sample Results

Deep Learning Model	Weights	Precision	Recall	f1-score	Accuracy	Standard Error of Accuracy
InceptionResNetV2	**ImageNet**	0.906	0.902	**0.902**	**0.902**	0.001
	None	0.841	0.839	0.839	0.839	0.003

Rai, T., et al. "Can ImageNet feature maps be applied to small histopathological datasets for the classification of breast cancer metastatic tissue in whole slide images?." *Medical Imaging 2019: Digital Pathology*. Vol. 10956. International Society for Optics and Photonics, 2019.

Conclusion and further work

- Deep learning models are shown to be effective and promising in digital pathology
- Development and validation of emerging technologies to aid cancer diagnosis
  - Improving diagnostic accuracy for tumour detection and grading
- Streamline the work load for pathologists
  - Focus on the complex border line cases hard to classify
  - Reducing exposure to simply diagnostic examples
- Aid veterinary pathologists, oncologists and surgeons to make better clinical decisions with regard to the diagnosis
  - Providing better patient care and outcomes


Acknowledgments
zoetis
NPL
UNIVERSITY OF SURREY
Colorado State University
fitzpatrickreferrals
UF
UNIVERSITY of FLORIDA

4. Data management plan: hand-on exercise

"**Data management**" plans are a catchphrase of funding agencies these days. However, they are not just useful for academic projects: in almost any group project, data is what we create and controls the collaboration. A data management plan also helps here, so that people think about how data is stored and shared first. In how many projects, have you ran out of energy because the necessary data/documents are spread all over the place?

This is a simple data management plan. It should be simple enough to fill out for any project. Be efficient: only fill out what is needed, not everything applies to every project.

4.1. Introduction

A Data Management Plan (DMP) outlines how data will be created, managed, shared and preserved. It helps you save time and effort, check that necessary support is in place, enables sound decisions, demonstrates awareness of good practice and reassures funders that the proposal is in line with their data policy.

- Read the research project scenario;
- Identify the potential data management issues on page 7 and begin to develop the DMP;
- Use the checklist to help you complete the DMP;
- Complete the H2020 DMP template.

4.2. Research project scenario

Roles and responsibilities

You are a Principal Investigator (PI) on this project and plan to lead a 5 years research project involving three Universities, along with international collaborators.

The program structure

The program of work is divided into several streams and will study virus like particles (VLP) and inorganic engineered nanomaterials automatic detection techniques, as well as the development of identification and classification methods, involving both chemical (composition, mass and number concentration) and physical information (e.g. size, shape, aggregation) using electron microscopy. The first four streams will address four key questions in classification and identification. The fifth stream will study the correlation between the different detection techniques and the four previous stream.

Expected data

The project will involve a number of analytical and research methods and methodologies that will be used to collect and analyze data. These will consist of quantification measures, qualification and classification models and the development of computational comparison models of the different automatic detection techniques. 150 micrographs are generated and analyzed for each of the seven different nanomaterials for a total of 1050 micrographs. The precision and accuracy of 15 different thresholding algorithms used for automatic nanoparticles detection in the micrographs analyzed & compared.

Period of data retention

This project proposal is part of a larger project to collect and analyze automatic nanoparticles detection, analysis and classification methods, in micrographs, using transmission electron microscopy. As such, it is necessary to set an embargo period of one (1) year in order to allow the PI first use rights. After the embargo period, the public-use version of the data will be available via Archivematica, where the data will be preserved, enhanced with standardized descriptive DDI (Data Documentation Initiative) metadata, and made available online indefinitely. Data and analyses need to be shared between all organizations throughout the five-year program.

Licensing

At the end of the embargo period, data will be made accessible under the CC-BY-NC-SA. Existing data provided by project partners will be used in some work packages, as well as data that is available under a Creative Commons CC-BY-NC-SA (Attribution-Noncommercial-Share Alike) License.

Data Format

All data collected during the course of this project will be cleaned, processed, and analyzed using the statistical software package and will consist of an SPSS file. This file, along with metadata (codebook, description of methodology, Etc.) included for archiving with the final product, should make it easy for researchers and the public to interpret and use the data. Existing data will be re-analyzed and the new analyses stored as Excel and SPSS (IBM software platform for advances statistical analysis) datasheets in the project archives. All data will be analyzed, stored and documented, and will include MS Excel, SPSS, and R Studio files. All data will be collected where possible in a paper-free method directly onto laptops or hand-held devices. The Transmission Electron Microscope automatically generates micrographs using the TIFF file format.

Naming convention

Micrographs will be named using the following standard convention: [A-E]'[NUMBER][K-Q][NUMBER]

- File name starts with a capital letter representing the year of the project where A represent the first year of the project.
- The first letter will be followed by the **prime** sign if the negative staining method is used
- This will be followed by a number indicating the grid number
- Last but not least, a capital letter raging from **K** to **Q** will indicate the nanomaterial used
- The naming ends with a number indicating the used vial

Data sharing

Data will be made available to the public following a one-year embargo period. PI will be collaborating with the Archivematica for long-term preservation description, backup, and dissemination of the data. This dataset will be made openly available online through the Archivematica website at https://www.archivematica.org/en/

Data Storage and preservation of access

Micrographs will be recorded using the hard drive of the electron microscope and then transferred to the laboratory storage server for local preservation. Paper copies and manuals generated from the research are digitized and transferred to the laboratory storage server while hard copies will be stored in the laboratory file cabinet. All long-term maintenance, curation, and archiving of the data will be done in partnership Archivematica. Archivematica processing and preservation procedures are based on the Reference Model for an Open Archival Information System, and include robust migration plans, technology monitoring, and preservation strategies. Beyond the life of the project, data will be preserved in partnerships with Archivematica indefinitely. All metadata will be available, minus any data that contains personally identifying information. Metadata submitted to accompany this data set will include a description of the study methodology and a codebook to assist in the interpretation and use of the data.

Data security and Ethical concerns

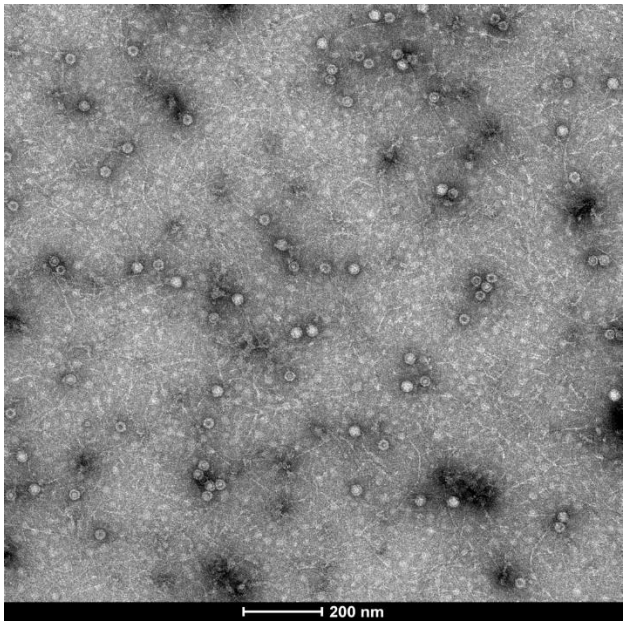
Before dissemination and use, data will be scrubbed of personally identifying information to the extent possible. Where not possible, datasets that contain personally identifiable information will be secured using the following measures: file encryption, password protection, and access restrictions. Aside from the protection of these datasets, no other ethical or privacy issues have been identified.

Intellectual property and copyrights

Intellectual property and copyright are held across the partner institutions.

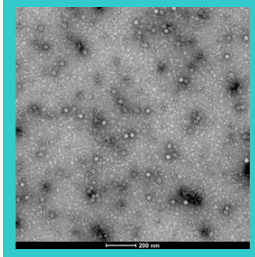
Additional possible data management requirements

4.3. Data and Metadata Example



Data

Exif Info: H'124N4.tif



File

Filename
H'124N4.tif

File Size
17 MB

File Type
TIFF

File Type Extension
tif

MIME Type
image/tiff

Exif Byte Order
Little-endian (Intel, II)

EXIF

Image Width
4096

Image Height
4224

Bits Per Sample
8

Compression
Uncompressed

Photometric Interpretation
BlackIsZero

Strip Offsets
162

Rows Per Strip
4224

Strip Byte Counts
17301504

X Resolution
26649426.74

Y Resolution
26649426.74

Resolution Unit
cm

Composite

Image Size
4096x4224

Megapixels
17.3

https://exifinfo.org/detail/MSZA6Ue6zyCC87phBS-SzA

Metadata

- Contextual information for data is called metadata — literally data about data
- Helpful for the reproduction of your work
- Should clearly and explicitly include the identifier of the data they describe
- Like data, metadata are registered or indexed in a searchable resource
- Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.
- Since storing the metadata generally is much easier and cheaper than the data itself, finding the institution researcher or any needed information to track down the original information Is still possible using only metadata

So as a rule of thumb, you should never say ‘this metadata isn’t useful’; be generous and provide it anyway!

```
<Root>
  <Data>
    <Label>Microscope</Label>
    <Value>Tecnai 12 D1095 BioTwin</Value>
    <Unit/>
  </Data>
  <Data>
    <Label>User</Label>
    <Value>SUPERVISOR</Value>
    <Unit/>
  </Data>
  <Data>
    <Label>Gun type</Label>
    <Value>LaB6</Value>
    <Unit/>
  </Data>
  <Data>
    <Label>High tension</Label>
    <Value>120</Value>
    <Unit>kV</Unit>
  </Data>
  <Data>
    <Label>Wehnelt index</Label>
    <Value>4</Value>
    <Unit/>
  </Data>
  <Data>
    <Label>Emission</Label>
    <Value>3.42</Value>
    <Unit>uA</Unit>
  </Data>
  <Data>
    <Label>Mode</Label>
    <Value> TEM uP SA Zoom Image</Value>
    <Unit/>
  </Data>
  <Data>
    <Label>Defocus</Label>
    <Value>-0.281</Value>
    <Unit>um</Unit>
  </Data>
  <Data>
    <Label>Magnification</Label>
    <Value>30000</Value>
    <Unit>x</Unit>
  </Data>
  <Data>
    <Label>Spot size</Label>
    <Value>1</Value>
    <Unit/>
  </Data>
  <Data>
    <Label>Intensity</Label>
    <Value>39.623</Value>
    <Unit>%</Unit>
  </Data>
  <Data>
    <Label>Objective lens</Label>
    <Value>90.894</Value>
    <Unit>%</Unit>
  </Data>
  <Data>
    <Label>Diffraction lens</Label>
    <Value>61.436</Value>
    <Unit>%</Unit>
  </Data>
  <Data>
    <Label>Stage X</Label>
    <Value>337.601</Value>
    <Unit>um</Unit>
  </Data>
  <Data>
    <Label>Stage Y</Label>
    <Value>599.951</Value>
    <Unit>um</Unit>
  </Data>
  <Data>
    <Label>Stage Z</Label>
    <Value>62.547</Value>
    <Unit>um</Unit>
  </Data>
  <Data>
    <Label>Stage A</Label>
    <Value>0.07</Value>
    <Unit>deg</Unit>
  </Data>
  <Data>
    <Label>Stage B</Label>
    <Value>0.00</Value>
    <Unit>deg</Unit>
  </Data>
</Root>
```

4.4. Summary of issues for DMP sample

List of issues	DMP template headings
• Is there evidence that secondary sources of data have been considered and evaluated? • Is there evidence presented that the project is not creating new data when there are existing resources that could be re-used? • If existing data are used, have issues such as copyright or IPR of such data been considered and possible copyright clearance obtained to be able to share data or data derived thereof?	Assessment of existing data
• Is the information on data to be produced adequate and realistic and according to the research and methodology proposed in the application? • Is there evidence that the plan covers all data that is planned to be generated from the research? • Is sufficient information given on how data will be collected and in which formats (eg Open Document Format, tab-delimited, Excel etc) data will be analyzed and stored, as well as an indication of how they will be documented?	Information on new data
• Is information given on procedures for quality assurance that will be carried out on the data collected? This could include methods for data validation or standards applied during data collection and data entry, codes of research practice adhered to, transcription templates used, etc. • Are no quality assurance procedures mentioned when there is a clear need from the proposed research that there should be? Please note that quality issues are to be addressed at the time of data collection, data entry, digitization or data checking.	Quality assurance of data
• Is the data back-up procedure described fit for purpose? eg considering back-up procedures for all institutions involved in research and considering back-up frequency • Are multiple media and multiple copies considered for back-up? • Are measures considered to check the usability of back-up copies? • Is information given on an institutional and/or local center back-up policy? • If sensitive data (i.e. detailed personal data) are collected, is there evidence that appropriate security measures in line with the Data Protection Act are considered when handling and storing data? e.g. encrypting data, anonymizing data, care when transmitting data • Is there evidence presented that proposed measures reflect existing best practices? • Are methods of version control described? (i.e. making sure that if the information in one file is altered, the related information in other files is also adopted, as well as keeping a track on a number of versions and their locations)	Backup and security of data
• Have all obstacles to sharing data been considered? • Have strategies been considered for dealing with these issues? For example by: ○ discussing data sharing and re-use and gaining specific consent to share research data ○ anonymizing data to remove personal information ○ regulating access to data ○ If there are ethical issues which may cause difficulties in data sharing, strategies for dealing with these issues should be discussed in the relevant section.	Expected difficulties in data sharing

○ If newly generated data cannot be shared, adequate justification should be given. It may be a case that parts of the data that are sensitive cannot be shared, but this should be considered critically and the plan should provide evidence that it has been assessed from all angles.	
• Is copyright of research data (both existing sources of data used or created) agreed or clarified, especially for collaborative research or if various sources of data are combined? • Are plans in place for copyright clearance for data sharing (if possible)?	Copyright/intellectual property right
• Have data management responsibilities been allocated to named individuals? • Is there evidence that data management will be followed throughout the course of the project? • Has consideration been given to the variety of data management tasks that may be required for the research? • For collaborative research, are data management responsibilities allocated at each partner organization (if needed for the research) or has the coordination of data management responsibilities across partners been considered?	Responsibilities
• Are the plans for preparing and documenting data for sharing and archiving with the funding agency are appropriate? • Is there evidence that data will be well documented during research to provide high-quality contextual information and/or structured metadata for secondary users? eg documenting the method of data collection, origin, circumstances, processing and analysis of data	Preparation of data for sharing and archiving

4.5. Summary of issues for DMP

List of issues	DMP template headings
	Assessment of existing data
	Information on new data
	Quality assurance of data
	Backup and security of data
	Expected difficulties in data sharing
	Copyright/intellectual property right
	Responsibilities
	Preparation of data for sharing and archiving

4.6. Data Management Plan Checklist Sample

DCC Checklist	DCC Guidance and questions to consider
Administrative Data	
ID	A pertinent ID as determined by the funder and/or institution.
Funder	State research funder if relevant
Grant Reference Number	Enter grant reference number if applicable [POST-AWARD DMPs ONLY]
Project Name	If applying for funding, state the name exactly as in the grant proposal.
Project Description	**Questions to consider:** - What is the nature of your research project? - What research questions are you addressing? - For what purpose are the data being collected or created? **Guidance:** Briefly summarise the type of study (or studies) to help others understand the purposes for which the data are being collected or created.
PI / Researcher	Name of Principal Investigator(s) or main researcher(s) on the project.
PI / Researcher ID	E.g ORCID http://orcid.org/
Project Data Contact	Name (if different to above), telephone and email contact details
Date of First Version	Date the first version of the DMP was completed
Date of Last Update	Date the DMP was last changed
Related Policies	**Questions to consider:** - Are there any existing procedures that you will base your approach on? - Does your department/group have data management guidelines? - Does your institution have a data protection or security policy that you will follow? - Does your institution have a Research Data Management (RDM) policy? - Does your funder have a Research Data Management policy? - Are there any formal standards that you will adopt? **Guidance:** List any other relevant funder, institutional, departmental or group policies on data management, data sharing and data security. Some of the information you give in the remainder of the DMP will be determined by the content of other policies. If so, point/link to them here.
Data Collection	
What data will you collect or create?	**Questions to consider:** - What type, format and volume of data? - Do your chosen formats and software enable sharing and long-term access to the data? - Are there any existing data that you can reuse? **Guidance:** Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.
How will the data be collected or created?	**Questions to Consider:** - What standards or methodologies will you use? - How will you structure and name your folders and files? - How will you handle versioning? - What quality assurance processes will you adopt? **Guidance:** Outline how the data will be collected/created and which community data standards (if any) will be used. Consider how the data will be organised during the project, mentioning

	for example naming conventions, version control and folder structures. Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeat samples or measurements, standardised data capture or recording, data entry validation, peer review of data or representation with controlled vocabularies.
Documentation and Metadata	
What documentation and metadata will accompany the data?	**Questions to consider:** - What information is needed for the data to be to be read and interpreted in the future? - How will you capture / create this documentation and metadata? - What metadata standards will you use and why? **Guidance:** Describe the types of documentation that will accompany the data to help secondary users to understand and reuse it. This should at least include basic details that will help people to find the data, including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed. Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data. Consider how you will capture this information and where it will be recorded. Wherever possible you should identify and use existing community standards.
Ethics and Legal Compliance	
How will you manage any ethical issues?	**Questions to consider:** - Have you gained consent for data preservation and sharing? - How will you protect the identity of participants if required? e.g. via anonymisation - How will sensitive data be handled to ensure it is stored and transferred securely? **Guidance:** Ethical issues affect how you store data, who can see/use it and how long it is kept. Managing ethical concerns may include: anonymisation of data; referral to departmental or institutional ethics committees; and formal consent agreements. You should show that you are aware of any issues and have planned accordingly. If you are carrying out research involving human participants, you must also ensure that consent is requested to allow data to be shared and reused.
How will you manage copyright and Intellectual Property Rights (IPR) issues?	**Questions to consider:** - Who owns the data? - How will the data be licensed for reuse? - Are there any restrictions on the reuse of third-party data? - Will data sharing be postponed / restricted e.g. to publish or seek patents? **Guidance:** State who will own the copyright and IPR of any data that you will collect or create, along with the licence(s) for its use and reuse. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. Consider any relevant funder, institutional, departmental or group policies on copyright or IPR. Also consider permissions to reuse third-party data and any restrictions needed on data sharing.
Storage and Backup	
How will the data be stored and backed up during the research?	**Questions to consider:** - Do you have sufficient storage or will you need to include charges for additional services? - How will the data be backed up? - Who will be responsible for backup and recovery? - How will the data be recovered in the event of an incident? **Guidance:** State how often the data will be backed up and to which locations. How many copies are being made? Storing data on laptops, computer hard drives or external storage devices alone is very risky. The use of robust, managed storage provided by university IT teams is preferable. Similarly, it is normally better to use automatic backup services provided by IT Services than rely on manual processes. If you choose to use a third-party service, you

	should ensure that this does not conflict with any funder, institutional, departmental or group policies, for example in terms of the legal jurisdiction in which data are held or the protection of sensitive data.
How will you manage access and security?	**Questions to consider:** - What are the risks to data security and how will these be managed? - How will you control access to keep the data secure? - How will you ensure that collaborators can access your data securely? - If creating or collecting data in the field how will you ensure its safe transfer into your main secured systems? **Guidance:** If your data is confidential (e.g. personal data not already in the public domain, confidential information or trade secrets), you should outline any appropriate security measures and note any formal standards that you will comply with e.g. ISO 27001.
Selection and Preservation	
Which data should be retained, shared, and/or preserved?	**Questions to consider:** - What data must be retained/destroyed for contractual, legal, or regulatory purposes? - How will you decide what other data to keep? - What are the foreseeable research uses for the data? - How long will the data be retained and preserved? - **Guidance:** Consider how the data may be reused e.g. to validate your research findings, conduct new studies, or for teaching. Decide which data to keep and for how long. This could be based on any obligations to retain certain data, the potential reuse value, what is economically viable to keep, and any additional effort required to prepare the data for data sharing and preservation. Remember to consider any additional effort required to prepare the data for sharing and preservation, such as changing file formats.
What is the long-term preservation plan for the dataset?	**Questions to consider:** - Where e.g. in which repository or archive will the data be held? - What costs if any will your selected data repository or archive charge? - Have you costed in time and effort to prepare the data for sharing / preservation? **Guidance:** Consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.
Data Sharing	
How will you share the data?	**Questions to consider:** - How will potential users find out about your data? - With whom will you share the data, and under what conditions? - Will you share data via a repository, handle requests directly or use another mechanism? - When will you make the data available? - Will you pursue getting a persistent identifier for your data? **Guidance:** Consider where, how, and to whom data with acknowledged long-term value should be made available. The methods used to share data will be dependent on a number of factors such as the type, size, complexity and sensitivity of data. If possible, mention earlier examples to show a track record of effective data sharing. Consider how people might acknowledge the reuse of your data.
Are any restrictions on data sharing required?	**Questions to consider:** - What action will you take to overcome or minimise restrictions? - For how long do you need exclusive use of the data and why? - Will a data sharing agreement (or equivalent) be required? **Guidance:** Outline any expected difficulties in sharing data with acknowledged long-term value,

	along with causes and possible measures to overcome these. Restrictions may be due to confidentiality, lack of consent agreements or IPR, for example. Consider whether a non-disclosure agreement would give sufficient protection for confidential data.
Responsibilities and Resources	
Who will be responsible for data management?	**Questions to consider:** - Who is responsible for implementing the DMP, and ensuring it is reviewed and revised? - Who will be responsible for each data management activity? - How will responsibilities be split across partner sites in collaborative research projects? - Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between partners? **Guidance:** Outline the roles and responsibilities for all activities e.g. data capture, metadata production, data quality, storage and backup, data archiving & data sharing. Consider who will be responsible for ensuring relevant policies will be respected. Individuals should be named where possible.
What resources will you require to deliver your plan?	**Questions to consider:** - Is additional specialist expertise (or training for existing staff) required? - Do you require hardware or software which is additional or exceptional to existing institutional provision? - Will charges be applied by data repositories? **Guidance:** Carefully consider any resources needed to deliver the plan, e.g. software, hardware, technical expertise, etc. Where dedicated resources are needed, these should be outlined and justified.

4.7. Data Management Plan Checklist

Administrative Data	
ID	
Funder	
Grant Reference Number	
Project Name	
Project Description	
PI / Researcher	
PI / Researcher ID	
Project Data Contract	
Data of First Version	
Date of Last Version	
Related Policies	
Data Collection	
What data will you collect or create	
How will the data be collected or created?	
Documentation and Metadata	
What documentation and metadata will accompany the data?	
Ethics and Legal Compliance	
How will you manage any ethical issues?	
How will you manage copyright and intellectual property rights (IPR) issues?	
Storage and Backup	
How will the data be stored and backed up during the research?	
How will you manage access and security	
Selection and preservation	
Which data should be retained, shared, and/or preserved?	
What is the long-term preservation plan for the sataset?	
Data Sharing	
How will you share the data?	
Are any restriction on data sharing required?	
Responsibilities and Resources	
Who will be responsible for data management?	
What resources will you require to deliver your plan?	

4.8. Horizon 2020 FAIR DMP

Introduction

This Horizon 2020 FAIR DMP template has been designed to be applicable to any Horizon 2020 project that produces, collects or processes research data. You should develop a single DMP for your project to cover its overall approach. However, where there are specific issues for individual datasets (e.g. regarding openness), you should clearly spell this out.

FAIR data management

In general terms, your research data should be 'FAIR', that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation solution.

This template is not intended as a strict technical implementation of the FAIR principles, it is rather inspired by FAIR as a general concept.

More information about FAIR:

FAIR data principles (FORCE11 discussion forum)

FAIR principles (article in Nature)

Structure of the template

The template is a set of questions that you should answer with a level of detail appropriate to the project. It is not required to provide detailed answers to all the questions in the first version of the DMP that needs to be submitted by month 6 of the project. Rather, the DMP is intended to be a living document in which information can be made available on a finer level of granularity through updates as the implementation of the project progresses and when significant changes occur. Therefore, DMPs should have a clear version number and include a timetable for updates. As a minimum, the DMP should be updated in the context of the periodic evaluation/assessment of the project. If there are no other periodic reviews envisaged within the grant agreement, an update needs to be made in time for the final review at the latest.

In the following the main sections to be covered by the DMP are outlined. At the end of the document, Table 1 contains a summary of these elements in bullet form.

This template itself may be updated as the policy evolves.

4.9. H2020 Template Sample

DMP component	Issues to be addressed
1. Data summary	• State the purpose of the data collection/generation • Explain the relation to the objectives of the project • Specify the types and formats of data generated/collected • Specify if existing data is being re-used (if any) • Specify the origin of the data • State the expected size of the data (if known) • Outline the data utility: to whom will it be useful
2. FAIR Data 2.1. Making data findable, including provisions for metadata	• Outline the discoverability of data (metadata provision) • Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers? • Outline naming conventions used • Outline the approach towards search keyword • Outline the approach for clear versioning • Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how
2.2. Making data openly accessible	• Specify which data will be made openly available? If some data is kept closed provide rationale for doing so • Specify how the data will be made available • Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)? • Specify where the data and associated metadata, documentation and code are deposited • Specify how access will be provided in case there are any restrictions
2.3. Making data interoperable	• Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability. • Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?
2.4. Increase data re-use (through clarifying licenses)	• Specify how the data will be licenced to permit the widest reuse possible • Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed • Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why • Describe data quality assurance processes • Specify the length of time for which the data will remain re-usable
3. Allocation of resources	• Estimate the costs for making your data FAIR. Describe how you intend to cover these costs • Clearly identify responsibilities for data management in your project

	• Describe costs and potential value of long term preservation
4. Data security	• Address data recovery as well as secure storage and transfer of sensitive data
5. Ethical aspects	• To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former
6. Other	• Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

4.10. H2020 Template

DMP component	*Issues to be addressed*
1. Data summary	
2. FAIR Data 2.1. Making data findable, including provisions for metadata	
2.2. Making data openly accessible	
2.3. Making data interoperable	
2.4. Increase data re-use (through clarifying licences)	
3. Allocation of resources	
4. Data security	
5. Ethical aspects	
6. Other	

4.11. Example of DMP

Data management plan for:

Type(s) of data: What particular type of data is this plan covering, and what is the data used for? You can multiple forms for different types of data (e.g. data, software, registry data)

Non-technical

SOURCE: From where do you get data (new data collected, reused, or other)? What are the IPR licenses? Ethical considerations?

DOCUMENTATION: How do you document the data? Is the documentation sufficient to reproduce the data from the inputs, and your work from the data? How will you make it so that you can understand the data in 10 years? Processing scripts and code can be part of the documentation, too (and they should be documented). At minimum, docs should be in a README.txt file along with the data.

SOFTWARE: What software do you use to access data? What methods (e.g. statistical, algorithms) do you use? Can others reproduce these methods? Is software open/can others use it, now and in the future?

INTEROPERABILITY: How do you ensure that your data is machine-readable, compatible, and linkable with other data (your domain and others)? You should use standard terms and identifiers if available in your field (ontologies). Mention what standards within your field you will follow. If standards don't exist, what will you do instead?

SHARING: Who do you think could use your data (think outside of your domain and of unexpected uses, too)? What will be shared? What license and level of sharing will you use (not shared, by request, or open and public)? Any privacy concerns? Will there be an embargo period?

COLLABORATION: How will you allow collaboration during processing and analysis? Is collaboration restricted to people in your institution, or can others collaborate as well? Can someone find, and send improvements back without asking first?

Technical

INPUTS: In what formats do you receive or initially store the data? What's the size? How do you receive it securely? Any immediate processing? Does any raw data go straight to archival in addition to final data?

PROCESSING: How do you do preprocessing? Is this preprocessing reproducible? How do you validate the quality of the data? How do you name files so that they make sense later? How will data be versioned? What are the costs for all data management and how are they covered?

STORAGE: Where will data be stored? Who manages this?

FORMATS: What are your working data formats? Are they standard and open? Ideal formats are open, standard, and have multiple programs that can read them. Will formats be usable in the future?

SECURITY: How will data be kept secure? Consider: 1) confidentiality, 2) availability (backups), 3) integrity (how do you prevent unnoticed corruption, for example someone editing a master file)?

REPOSITORIES: What repositories are used to share the data? If not in a repository, how will people find the data? In what formats do you share the data? Will people be able to find the data? Will there be persistent identifiers?

END OF LIFE: How do you decide what to delete, save, and migrate? Must any data be deleted? Is the deleted data reproducible? How do you prevent bit rot? (The best solution is to put enough to be reproduced on an archive)

General instructions/advice: Make a copy, remove instructions and fill in your solutions. Fill out as concisely as possible - shorter and more standardized is always better. Where relevant, include names of who is responsible.

Data management plan: personal recipe collection

Type(s) of data: Recipes for cooking. I want to be able to store them electronically, and also want to be able to share them openly. I care about archival, and someday I might want to be able to automatically process them.

Non-technical

SOURCE: Data comes from friends who provide resources or online sites whose recipes we want to save. I will only record data which can be relicensed under CC-BY, and any external people who add recipes will agree to allow me to license and provide access at will with any open license. I won't add any recipes with ethical issues (for example, possibly medically dangerous recipes)

DOCUMENTATION: Recipes are currently collected in natural language and should be self-describing. Each recipe must at least have the metadata of the source, source license, author, and date. There will be a template which has the necessary metadata, so that it is easy to document things. We don't currently envision automatic processing, but if we do we'll make sure to document that

SOFTWARE: All recipes are currently read and processed only by humans, no special software is needed.

INTEROPERABILITY: There is a lot of possibility for confusion of food products, especially since I will be sharing to people with different levels of cooking knowledge and languages. For minor additives, we will cite the E-number. For other food products, we will try to also provide a scientific name of the food item or en:Wikipedia link for what it uses, with description accurate within wikipedia. However, none of these make the data automatically linkable. A BBC food ontology exists, which I will investigate sometime. It would be neat to make recipes automatically processable, but I'll have to decide if the effort is worth it first.

SHARING: I don't expect there to be a big audience for these recipes, but I want to be able to share them if desired. The whole folder will be public, though not really advertised so I don't expect many visitors. I want these recipes to be open (thus the concern about licensing at the beginning), and they will be shared under a sharealike-type license.

COLLABORATION: Google Drive provides fine-grained collaboration posibilities: we can share read-only and read-write, and also restrict sharing to certain recipes or the whole folder. This is easy to use, so everyone can contribute. Contributors have to agree to licensing terms.

Technical

INPUTS: Recipes may be written down by friends when visiting, or I may copy from a source online. Generally, I would first write it down on physical paper (even if it is already online), revise/adjust, then enter it electronically. This allows me to be sure that I have a license to redistribute the recipe.

PROCESSING: Once recipes are written on paper, they are used for the first trials. At this point, any notes will be written directly on the paper. This is not reproduciable, but that is OK because we consider cooking to be as much of a creative art as a science. After a few trials, they are added to the permanent storage location.

STORAGE: Recipes are stored in Google Drive. We make one folder and each recipe (or set of related recipes) have one document. I am the direct manager.

FORMATS: The storage format is the native Google Docs format. This allows good collaboration and ediing, and can be exported as different open formats. Downloads are made in zip+odf, which are both open source formats.

SECURITY: I'm not really worried about confidentiality (we won't put anything secret in this file), but if I'm using Google Drive I could lose my data at any time since it's a free service. I'll do a backup periodically (download zip file). The change and risk of corruption is low, and it should be detectable/fixable because Google Docs has a history feature (assuming I notice).

REPOSITORIES: We haven't identified an ideal repository yet. I can make our Google Drive folder directly readable by the whole world: this is useful but not long-term. I can archive an export to a long-term repository, but this is not usable for day to day work.

END OF LIFE: The intermediate paper copies will be recycled once recipes are added to Google Drive. We don't plan on deleting from Google Drive, if something gets deprecated I'lll move it to an archive folder. Currently, I assume Google Drive keeps things available long enough, but this will be reevaluated as needed.

Data management plan:

Type/purpose/owner of data:

Non-technical

SOURCE:	**DOCUMENTATION:**	**SOFTWARE:**	**INTEROPERABILITY:**	**SHARING:**
	COLLABORATION:			

Technical

INPUTS:	**PROCESSING:**	**STORAGE:**	**REPOSITORIES:**	**END OF LIFE:**
		FORMATS:		
		SECURITY:		